# Numerical simulation of one-dimensional transient vertical flow in variably saturated soils

Daniel Caviedes Voullième

*Máster en Mecánica Aplicada*
*Programa Oficial de Posgrado en Ingeniería Mecánica y de Materiales*

September 2010

Supervisor: Dr. Pilar García Navarro

Term 2009-2010

Centro Politécnico Superior — Universidad de Zaragoza


# Numerical simulation of one-dimensional transient vertical flow in variably saturated soils

## Abstract

Water flow in variably saturated (saturated/unsaturated) soils is commonly modeled by means of Richards' equation. This equation has no general analytical solution and the use of numerical approximations is necessary. It can be presented in three physically equivalent forms which are based on different variables and show different mathematical properties. In this work, these forms are derived from the general mathematical model and analyzed from a numerical perspective in order to understand the interactions between the differential equations and the numerical methods required in each case.

The goals of this work are, on the one hand, to describe the physical and mathematical reasoning which leads to the formulation of the general mathematical model of flow in porous media, the discussion of the concepts and assumptions which allow to develop Richards' equation, and on the other hand, to establish the properties and limitations of several numerical schemes to approximate the solutions of flows in variably saturated soils.

The approach for the mathematical model is to average a microscopic, single-phase flow equation into a macroscopic scale which allows to describe porous media in a practical way, and to consider the necessary assumptions to state Richards' equation as a particular flow case. The porous media constitutive model completes the mathematical model. In this work the Mualem-van Genuchten model and some variants are included.

For the numerical model, several schemes are developed for the 1D Richards' equation in the vertical direction. Explicit and implicit centered finite difference schemes are used in this work. The key numerical aspects of interest are those of mass conservation, stability and efficiency. Another key aspect, which is not only numerical is that of continuity from unsaturated into saturated regimes. The constitutive models affect the numerical schemes and some issues arise because of the high non-linearity of the functions, in particular the hydraulic conductivity function. Appropriate discretization of hydraulic conductivity for estimation of flux between numerical cells is a sensible issue wich has been studied by many authors and is treated in this work. All of these issues are analyzed individually and as interrelated problems in the schemes.

Validation and test cases are presented and the response of the model to different problems and parameters is examined. From them, it is concluded that the explicit and the implicit schemes based on the mixed form of Richards' equation are better suited for unsaturated problems. For variably saturated problems, the implicit scheme based on the mixed form is the best choice, since the explicit model cannot solve saturation conditions. Conditional stability of the explicit model affects negatively its performance in certain cases, which also leads to the conclusion that the implicit scheme is more efficient and realiable.

# Simulación numérica unidimensional de flujos transitorios verticales en suelos con saturación variable

## Resumen

El flujo de agua en suelos con saturación variable (saturado/no saturado) es comunmente modelizado por medio de la ecuación de Richards. Dicha ecuación no tiene una solución análitica general, y por tanto es necesario el uso de aproximaciones numéricas. La ecuación puede presentarse en tres formas, las cuales son físicamente equivalentes, pero basadas en distintas variables, y que muestra comportamientos matemáticos distintos. En este trabajo, dichas formas se obtienen a partir del modelo matemático general, y se analizan desde la perspectiva numérica con el fin de comprender las interacciones entre las formas de la ecuación diferencial y los métodos numéricos aplicables en cada caso.

Los objetivos de este trabajo son, por una parte, describir el razonamiento físico y matemático que lleva a la formulación del modelo matemático general de flujo en medios porosos, la discusión de los conceptos y supuestos que permiten formular la ecuación de Richards, y por otra, estudiar las propiedades y la aplicabilidad de los métodos numéricos para su solución.

El enfoque utilizado para el modelo matemático es el de promediar una ecuación microscópica de una sola fase, a una escala macroscópica que permita describir el medio poroso de una forma práctica y, posteriormente, considerar los supuestos que permiten formular la ecuación de Richards como un caso particular de flujo. El modelo constitutivo del medio poroso completa dicho modelo matemático. En este trabajo se utiliza el modelo de Mualem-van Genuchten así como una de sus variantes.

Para el modelo numérico, varios esquemas numéricos fueron formulados para la ecuación de Richards unidimensional, en la dirección vertical. En este trabajo se utilizan esquemas explícitos e implicitos con diferencias finitas centradas. Los aspectos de interés desde la perspectiva numérica son la conservación de masa, estabilidad y eficiencia computacional. Adicionalmente, un tema que no es únicamente numérico es la continuidad y aplicabilidad tanto en la región saturada como en la región parcialmente saturada. Los modelos constitutivos de suelo llevan al modelo numérico una serie de dificultades por las características de las funciones no lineales, en particular la conductividad hidráulica. La discretización cuidadosa the la conductividad entre las celdas es un tema de importancia, el cual ha sido estudiado por muchos autores y se incluye también en este trabajo. Todos estos aspectos se analizan en sí mismos y como problemas interrelacionados dentro de los esquemas numéricos.

Se presentan pruebas de validación y casos test y se examina la respuesta de los modelos a distintos problemas y parámetros. De dichas pruebas se puede concluir que los esquemas explícitos e implícitos basados en la forma mixta de la ecuación de Richards son más apropiados para la solución de problemas no saturados. Para condiciones de saturación variable, el esquema implícito basado en la forma mixta es la mejor opción, dado que el modelo explícito no es capaz de resolver condiciones de saturación. La estabilidad condicionada del modelo explícito también afecta de forma negativa a su eficiencia computacional en algunos casos, lo cual apoya la conclusión general de que el esquema implícito es más robusto y eficiente.

# Contents

# List of Figures

# List of Tables

# Introduction

The dynamics of flow in porous media is a field of study which has many applications, ranging from groundwater flow and underground petroleum flow to porous filters and porous flow in biological tissues. The general framework and theory is able to encompass all of these cases and to provide a mathematical model to understand the driving forces and the properties of such flows. This model gives rise to complex differential equations which require numerical methods to approximate their solutions.

This work lies within the science of flow in porous media, with particular interest in water flow in variably saturated soils. Throughout this work the term *variably* saturated flow is preferred over *unsaturated* flow as saturated flows are also of interest, and the transition from one to the other is one of the key aspects in study (in literature this is also refered to as *saturated/unsaturated flow*).

The formulation of the mathemical model is the first step. The general flow equations are examined from a formal (and general) microscopic approach and averaged into a macroscopic scale which allows to describe porous media as a continuum in a manageable way. To formulate Richards' equation from such point, assumptions need to be considered in order to neglect terms [4] [5] [6]. This is not the only possible approach, as Richards' equation may be obtained from simpler continuum models based on a differential control volume and Darcy's Law [7] [19] [23]. The latter approach is much more intuitive and practical, however, it does not allow to clearly identify and understand under which conditions variably saturated flow is a particular case in fluid dynamics. The first approach is more complex, but also more formal and does allow to understand Richards' equation as a particular case of conservation equations, hence, it is preferred in this work.

The mathematical model is complete only when the porous media constitutive model is included. Several models exist for soils [34]. Some of the best known are Brooks-Corey [9], Mualem-van Genuchten [37] and Gardner-Russo [16] [32]. These models relate pressure, water content and hydraulic conductivity with soil properties such as pore and grain size distributions.

Richards' equation [30] is the most commonly used model for flow in variably saturated soils [23]. This equation has no general analytical solution and the use of numerical approximations is necessary. It can be presented in three physically equivalent forms which are based on different variables and show different mathematical properties. In this work, these forms are analyzed from a numerical perspective in order to understand the interactions between the differential equations and the numerical methods required in each case.

In consequence, several schemes have been developed by several authors for the 1D Richards' equation in the vertical direction which are examined in this work. The horizontal directions are similar but simpler because the gravitational term is not present. Explicit and implicit centered finite difference schemes are used in this work. The key numerical aspects of interest

are those of mass conservation, stability and efficiency. Another key aspect which is not only numerical is that of continuity from unsaturated into saturated regimes. Authors such as Celia et al. [13], Rathfelder and Abriola [29], Phoon et al. [28] and Huang et al. [22] have delved into these aspects. Numerical issues arise because of the high non-linearity of the hydraulic conductivity function of the soil model. Appropriate estimation of hydraulic conductivity for computing flow between numerical cells is a sensible issue, and has been studied by many authors [2] [3] [8] [10] [14] [17] [20] [21] [27] [31] [35] [36] [40] [42].

All these aspects are included in the schemes presented in this work and are analyzed both as issues in their own right and as interrelated problems. In some cases one of the aforementioned issues may be much more significant than others, but in the general vision all of these aspects interact in complex ways and allow to conclude which combinations might best suit variably saturated flow problems.

The goals of this work are to describe the physical and mathematical reasoning which leads to the formulation of the general mathematical model of flow in porous media, the discussion of the concepts and assumptions which allow to develop Richards' equation, as well as describing the properties and establishing the range and limitations of different numerical schemes to approximate the solutions of flows in variably saturated soils.

This research lays some foundations for further study of the numerical solution of Richards' equation in 2D and 3D domains, and interactions with surface flow as well as substance transport in variably saturated media. In the long term, this work will be the basis upon which a 3D groundwater model will be coupled with a 2D surface model in order to study flow phenomena in rivers, channels, etc.

# Chapter 1

# Mathematical model and governing equations

The fundamental physical phenomena which govern flow in porous media are the same as in other branches of Fluid Dynamics when observed at a microscopic level, i.e., within a single fluid phase which is bounded by other fluid or solid phases. However, the microscopic scale is very difficult, if not impossible, to describe. The geometry is immensely complex, state variables are not easily defined and boundary conditions are quite difficult to formulate. Hence, it is necessary to take the analysis to a macroscopic scale, where variables and properties are averaged within a volume of appropriate size and characteristics: a Representative Elementary Volume (REV). The philosophy and reasoning within the following sections is based on the work by Bear, Bachmat and Verruijt [4] [5] [6].

Consider a representative volume of porous medium $U_o$, filled with a wetting phase $\alpha$ and any other phases $\beta$ as shown in figure 1.1. Let $E$ be an extensive property of phase $\alpha$. Let $e$ be the intensive property related to such property $E$. The size of the REV must be such that it is much smaller than the domain of the problem of interest, but large enough so that averages within it may smooth out inhomogeneities of the porous medium properties. Mathematically, within the REV, all quantities need to be continous and differentiable in time and space [24].

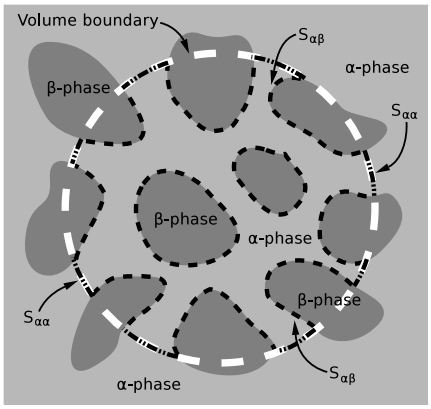



Figure 1.1: Representative Porous Volume

## 1.1 Microscopic equation

The microscopic approach intends to establish the general conservation equation of $E$ in the vicinity of a mathematical point. In the microscopic approach, only *one continous phase* $\alpha$ is considered so that the Navier-Stokes equations are valid, and could indeed be solved considering boundaries in the surface that contains such continous phase [6].

Let $\mathbf{V}$ be the velocity of the fluid, and $\mathbf{j}$ the difussive flux relative to the advective flux. Let $\rho$ be the density of phase $\alpha$ and $\Gamma^E$ the generation of $e$ within the volume $U_{o\alpha}$ of phase $\alpha$.

Thus, the rate of change of $e$ in the neighborhood of a point is given by

$$\frac{\partial e_\alpha}{\partial t} = -\nabla(e_\alpha \mathbf{V}_\alpha + \mathbf{j}_\alpha) + \rho\Gamma_\alpha^E \tag{1.1}$$

## 1.2 Averaging rules

In order to transform equation (1.1) into a macroscopic equation which is valid in a volume $U_o$ which contains several phases, it is necessary to introduce some averaging definitions, to transform mathematical point properties into representative properties in a control volume. Consider two phases and volume $U_o = U_{o\alpha} + U_{o\beta}$ as in figure 1.1 for the following definitions.

### 1.2.1 Averaging definitions

Let $\overline{e_\alpha}$ be the volumetric phase average of $e_\alpha$ in volume $U_o$.

$$\overline{e_\alpha} = \frac{1}{U_o}\int_{U_{o\alpha}} e_\alpha \, dU_\alpha \tag{1.2}$$

Let $\overline{e_\alpha}^\alpha$ be the volumetric intrinsic phase average of density $e_\alpha$ in the $\alpha$-phase fraction $U_{o\alpha}$ of volume $U_o$.

$$\overline{e_\alpha}^\alpha = \frac{1}{U_{o\alpha}}\int_{U_{o\alpha}} e_\alpha \, dU_\alpha \tag{1.3}$$

By defining the volume fraction $\theta_\alpha = \dfrac{U_{o\alpha}}{U_o}$, both averages can be related:

$$\overline{e_\alpha} = \theta_\alpha \overline{e_\alpha}^\alpha \tag{1.4}$$

Let $\mathring{e}_\alpha$ be the deviation of $e_\alpha$ of a mathematical point from the intrinsic phase average:

$$\mathring{e}_\alpha = e_\alpha - \overline{e_\alpha}^\alpha \tag{1.5}$$

Let $\widetilde{e}^{\alpha\beta}$ be the average of $e$ over a $S_{\alpha\beta}$ surface:

$$\widetilde{e}^{\alpha\beta} = \frac{1}{S_{\alpha\beta}}\int_{S_{\alpha\beta}} e \, dS \tag{1.6}$$

Let $\Sigma_{\alpha\beta}$ be the specific area of $S_{\alpha\beta}$, this is $\Sigma_{\alpha\beta} = \frac{S_{\alpha\beta}}{U_o}$, so that

$$\widetilde{e}^{\alpha\beta}\Sigma_{\alpha\beta} = \frac{1}{U_o}\int_{S_{\alpha\beta}} e \, dS \tag{1.7}$$

A relevant property of the intrinsic phase average is that it is a linear operator. Then, any quantity $G$ satisfies,

$$\overline{G_1 + G_2}^\alpha = \overline{G_1}^\alpha + \overline{G_2}^\alpha \tag{1.8}$$

The intrinsic phase average of a product of $G$ defined, following (1.3) by

$$\overline{G_1 G_2}^\alpha = \frac{1}{U_{o\alpha}}\int_{U_{o\alpha}} G_1 G_2 \, dU_\alpha$$

Because of (1.5), and considering that the intrinsic average of the fluctuations is zero,

$$\overline{G_1 G_2}^{\alpha} = \overline{G_1}^{\alpha}\overline{G_2}^{\alpha} + \overline{\mathring{G}_1\mathring{G}_2}^{\alpha} \tag{1.9}$$

### 1.2.2 Average of the time derivative

Consider Reynolds theorem for the extensive property $E$, over a volume $U_{o\alpha}$ contained by a surface $S_\alpha = S_{\alpha\alpha} + S_{\alpha\beta}$ containing such volume with normal outwards pointing vector $\hat{\mathbf{n}}$. $S_{o\alpha}$ can be considered as the sum of the contact surface between phases $\alpha$ and $\beta$ and the surface which separates phase $\alpha$ within the volume $U_o$ and the outside of such volume ($S_{\alpha\alpha}$).

$$\frac{D_E}{Dt}\int\limits_{U(t)} e\, dU = \int\limits_{U(t)} \frac{\partial e}{\partial t}\, dU + \int\limits_{S(t)} e\mathbf{V}^E\hat{\mathbf{n}}\, dS$$

where $\dfrac{D_E G}{Dt} = \dfrac{\partial G}{\partial t} + \mathbf{V}^E\nabla G$ is the material derivative with respect to an observer moving with $S(t)$.

**Assumption 1.2.1.** $U_{o\alpha}$ is assumed a *material volume* with respect to $E$, hence $S_{o\alpha}$ is a *material surface* which implies that $\mathbf{V}^E = \mathbf{u}$ for surface $S_{\alpha\beta}$ where $\mathbf{u}$ is the velocity at which $S_{\alpha\beta}$ is being displaced.

$$\frac{D_E}{Dt}\int\limits_{U_{o\alpha}(t)} e\, dU = \int\limits_{U_{o\alpha}(t)} \frac{\partial e}{\partial t}\, dU + \int\limits_{S_{\alpha\beta}(t)} e\mathbf{u}\hat{\mathbf{n}}\, dS + \int\limits_{S_{\alpha\alpha}(t)} e\mathbf{V}^E\hat{\mathbf{n}}\, dS \tag{1.10}$$

On the other hand, by considering the entire volume $U_o$, it is possible to express the material rate of change of the extensive quantity $E$ which only exists within phase $\alpha$. To do this, consider the characteristic function $\gamma_\alpha$ for phase $\alpha$:

$$\gamma_\alpha = \begin{cases} 1 \text{ , for points within } U_{o\alpha} \\ 0 \text{ , for points outside } U_{o\alpha} \end{cases} \tag{1.11}$$

Then, the material rate of change of $E$ within the entire volume is

$$\frac{D_E}{Dt}\int\limits_{U_o} \gamma_\alpha e\, dU = \frac{\partial}{\partial t}\int\limits_{U_o} \gamma_\alpha e\, dU + \int\limits_{S_o} \gamma_\alpha e\mathbf{V}^E\hat{\mathbf{n}}\, dS$$

which, when evaluating $\gamma_\alpha$ yields

$$\frac{D_E}{Dt}\int\limits_{U_{o\alpha}} e\, dU = \frac{\partial}{\partial t}\int\limits_{U_{o\alpha}} e\, dU + \int\limits_{S_{\alpha\alpha}} e\mathbf{V}^E\hat{\mathbf{n}}\, dS \tag{1.12}$$

By equating (1.10) and (1.12) considering instantaneously all time dependent terms,

$$\underbrace{\frac{\partial}{\partial t}\int\limits_{U_{o\alpha}} e\, dU}_{①} = \underbrace{\int\limits_{U_{o\alpha}} \frac{\partial e}{\partial t}\, dU}_{②} + \int\limits_{S_\alpha\beta} e\mathbf{u}\hat{\mathbf{n}}\, dS$$

Note that terms ① and ② may be substituted by using equation (1.3)

$$\frac{\partial}{\partial t}\left(\overline{e_\alpha}^{\alpha} U_{o\alpha}\right) = U_{o\alpha}\overline{\frac{\partial e_\alpha}{\partial t}}^{\alpha} + \int\limits_{S_\alpha\beta} e\mathbf{u}\hat{\mathbf{n}}\, dS$$

By dividing the entire equation by $U_o$ (which is constant in time, hence can go into the derivatives), and recalling that $\theta_\alpha = \dfrac{U_{o\alpha}}{U_o}$ and rearranging, yields

$$\theta_\alpha \overline{\frac{\partial e_\alpha}{\partial t}}^{\alpha} = \frac{\partial}{\partial t}\left(\overline{e_\alpha}^{\alpha}\theta_\alpha\right) - \overbrace{e\mathbf{u}\hat{\mathbf{n}}}^{\alpha\beta}\Sigma_{\alpha\beta} \tag{1.13}$$

This equation relates the average of a time derivative of $e_\alpha$ to the time derivative of the average of $e_\alpha$.

### 1.2.3 Average of the spatial derivative

From Gauss' theorem,

$$\int_{U_{o\alpha}} \nabla G \, dU = \int_{S_{\alpha\alpha}} G\hat{\mathbf{n}} \, dS + \int_{S_{\alpha\beta}} G\hat{\mathbf{n}} \, dS \tag{1.14}$$

Consider the integral over $S_{\alpha\alpha}$, by means of the characteristic function $\gamma_\alpha$ shown in (1.11)

$$\int_{S_{\alpha\alpha}} G\hat{\mathbf{n}} \, dS = \int_{S_o} G\gamma_\alpha \hat{\mathbf{n}} \, dS$$

Using Gauss' theroem once again

$$\int_{S_{\alpha\alpha}} G\hat{\mathbf{n}} \, dS = \int_{U_o} \nabla\left(G\gamma_\alpha\right) \, dU$$

Because $U_o$ does not change in space, the order of integration and differentiation can be exchanged, and furthermore, evaluating $\gamma_\alpha$,

$$\int_{S_{\alpha\alpha}} G\hat{\mathbf{n}} \, dS = \nabla \int_{U_o} G\gamma_\alpha \, dU = \nabla \int_{U_{o\alpha}} G \, dU$$

Substituting in (1.14),

$$\int_{U_{o\alpha}} \nabla G \, dU = \nabla \int_{U_{o\alpha}} G \, dU + \int_{S_{\alpha\beta}} G\hat{\mathbf{n}} \, dS$$

By means of (1.3) and (1.7)

$$U_{o\alpha}\overline{\nabla G}^{\alpha} = \nabla\left(U_{o\alpha}\overline{G}^{\alpha}\right) + U_o\overbrace{G\hat{\mathbf{n}}}^{\alpha\beta}\Sigma_{\alpha\beta}$$

Dividing the entire equation by $U_o$, and because $\theta_\alpha = \dfrac{U_{o\alpha}}{U_o}$ finally yields

$$\theta\overline{\nabla G}^{\alpha} = \nabla\left(\theta\overline{G}^{\alpha}\right) + \overbrace{G\hat{\mathbf{n}}}^{\alpha\beta}\Sigma_{\alpha\beta} \tag{1.15}$$

## 1.3 Macroscopic equation

Integrating equation (1.1) over the volume of phase $\alpha$, i.e. $U_{o\alpha}$, and dividing by the porous medium representative volume $U_o$, yields

$$\frac{1}{U_o}\int_{U_{o\alpha}} \frac{\partial e}{\partial t} = -\frac{1}{U_o}\int_{U_{o\alpha}} \nabla(e\mathbf{V} + \mathbf{j}) + \frac{1}{U_o}\int_{U_{o\alpha}} \rho\Gamma^E$$

By using equations (1.13) and (1.15),

$$\frac{\partial(\theta\overline{e}^{\alpha})}{\partial t} = -\nabla\left[\theta\Big(\underbrace{\overline{e}^{\alpha}\overline{\mathbf{V}}^{\alpha}}_{\text{Advective flux}} + \underbrace{\overline{\mathring{e}\mathring{\mathbf{V}}}^{\alpha}}_{\text{Dispersive flux}} + \underbrace{\overline{\mathbf{j}}^{\alpha}}_{\text{Diffusive flux}}\Big)\right] - \underbrace{\overbrace{[e\,(\mathbf{V}-\mathbf{u})+\mathbf{j})]\hat{\mathbf{n}}}^{\alpha\beta}\,\Sigma_{\alpha\beta}}_{\text{Surface flux}} + \underbrace{\theta\overline{\rho\Gamma^{E}}^{\alpha}}_{\text{Source/Sink}} \tag{1.16}$$

which is the general macroscopic conservation equation for any property $e$. Note that the equation includes advective, dispersive and diffusive fluxes, as well as flux through the interphase surface and a source/sink term.

## 1.4 Macroscopic mass balance equations

Consider mass $m$ as the extensive property $E$ in $\alpha$-phase of a single component. Hence, the intensive property $e$ becomes mass denstity $\rho$.

By taking equation (1.16), and $\Gamma^{E} = \Gamma^{m} = 0$ since mass is not generated within the volume, the general mass equation is obtained.

$$\frac{\partial(\theta\overline{\rho}^{\alpha})}{\partial t} = -\nabla\left[\theta\left(\overline{\rho}^{\alpha}\overline{\mathbf{V}}^{\alpha} + \overline{\mathring{\rho}\mathring{\mathbf{V}}}^{\alpha} + \overline{\mathbf{j}}^{\alpha}\right)\right] - \overbrace{[\rho\,(\mathbf{V}-\mathbf{u})+\mathbf{j}]\,\hat{\mathbf{n}}}^{\alpha\beta}\,\Sigma_{\alpha\beta} \tag{1.17}$$

By applying the intrinsic phase average of a product defined in equation (1.9) to the definition of the tensor quantity $\overline{\mathbf{j}}^{\alpha}$, together with the linear operator property shown in (1.8), it is possible to obtain after some manipulation,

$$\frac{\partial(\theta\overline{\rho}^{\alpha})}{\partial t} = -\nabla\left[\theta\left(\overline{\rho}^{\alpha}\overline{\mathbf{V}}^{\alpha} + \overline{\mathring{\rho}\mathring{\mathbf{V}}}^{\alpha}\right)\right] - \overbrace{\rho\,(\mathbf{V}-\mathbf{u})\,\hat{\mathbf{n}}}^{\alpha\beta}\,\Sigma_{\alpha\beta} \tag{1.18}$$

**Assumption 1.4.1.** There is no mass exchange between phases $\alpha$ and $\beta$. Hence, surface $S_{\alpha\beta}$ is a *material surface* respect to mass, which implies $\mathbf{V}-\mathbf{u}=0$.

$$\frac{\partial(\theta\overline{\rho}^{\alpha})}{\partial t} = -\nabla\left[\theta\left(\overline{\rho}^{\alpha}\overline{\mathbf{V}}^{\alpha} + \overline{\mathring{\rho}\mathring{\mathbf{V}}}^{\alpha}\right)\right] \tag{1.19}$$

**Assumption 1.4.2.** For each fluid phase, the sum of the dispersive and diffusive fluxes of the total mass is much smaller than the advective term.

Considering assumptions 1.4.1 and 1.4.2, equation (1.19) reduces to

$$\frac{\partial(\theta\overline{\rho}^{\alpha})}{\partial t} = -\nabla\left(\theta\overline{\rho}^{\alpha}\overline{\mathbf{V}}^{\alpha}\right) = -\nabla\left(\overline{\rho}^{\alpha}\mathbf{q}_{\alpha}\right) \tag{1.20}$$

where $\mathbf{q}_{\alpha} = \theta\overline{\mathbf{V}}^{\alpha}$ is the specific discharge of $\alpha$-phase. Equation (1.20) is a mass conservation equation with predominant advection and immiscible phases.

## 1.5 Mass balance in a non deformable, variably saturated porous medium

Consider a REV such as the one shown in figure 1.2. Let $U_o$ contain three phases only, one of which is a solid phase (s). Consider the other two as fluid phases: a wetting phase (w) and a non-wetting phase (n). A simple case of this is to imagine liquid water and air in soils.

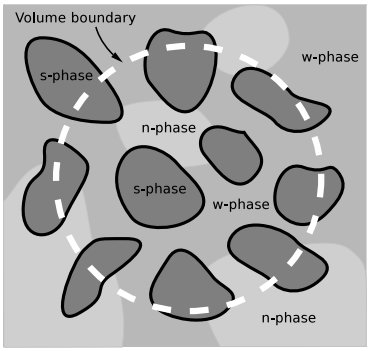



Figure 1.2: Three-phase representative volume

Let $S_\alpha$ be saturation $S_\alpha = \dfrac{\theta_\alpha}{\eta}$ where $\eta$ is porosity, and consider that the specific discharge is $\mathbf{q}_\alpha = \theta_\alpha \mathbf{V}_\alpha = \mathbf{q}_{r\alpha} + \theta_\alpha \mathbf{V}_s$, where $\mathbf{q}_{r\alpha}$ is the specific discharge of $\alpha$-phase relative to the (in the general case) moving solid with velocity $\mathbf{V}_s$.

Hence, the mass balance equation (1.20) of the wetting phase $\alpha = w$ can be written as

$$S_w\rho_w\frac{\partial\eta}{\partial t} + \eta\rho_w\frac{\partial S_w}{\partial t} + \eta S_w\frac{\partial\rho_w}{\partial t} = -S_w\mathbf{q}_{rw}\nabla\rho_w - \rho_w\mathbf{q}_{rw}\nabla S_w - S_w\rho_w\nabla\mathbf{q}_{rw} \tag{1.21}$$

In a similar way, the mass balance equation for the nonwetting phase is obtained by making $\alpha = n$.

$$S_n\rho_n\frac{\partial\eta}{\partial t} + \eta\rho_n\frac{\partial S_n}{\partial t} + \eta S_n\frac{\partial\rho_n}{\partial t} = -S_n\mathbf{q}_{rn}\nabla\rho_n - \rho_n\mathbf{q}_{rn}\nabla S_n - S_n\rho_n\nabla\mathbf{q}_{rn} \tag{1.22}$$

The equation for the solid phase is

$$\frac{1}{1-\eta}\frac{D_s}{Dt}(1-\eta) + \frac{1}{\rho_s}\frac{D_s\rho_s}{Dt} = -\nabla\mathbf{V}_s \tag{1.23}$$

where the material derivative $\dfrac{D_\alpha G}{Dt} = \dfrac{\partial G}{\partial t} + \mathbf{V}_\alpha\nabla G$ refers to velocity $\mathbf{V}_\alpha$.

By combining equation (1.21) and (1.22) with (1.23), the wetting phase mass balance equation is obtained,

$$\frac{\eta S_w}{\rho_w}\frac{D_w\rho_w}{Dt} + \frac{S_w}{1-\eta}\frac{D_s\eta}{Dt} + \eta\frac{D_sS_w}{Dt} - \frac{\eta S_w}{\rho_s}\frac{D_s\rho_s}{Dt} = -\nabla\mathbf{q}_{rw} \tag{1.24}$$

and the nonwetting phase

$$\frac{\eta S_w}{\rho_n}\frac{D_n\rho_n}{Dt} + \frac{S_n}{1-\eta}\frac{D_s\eta}{Dt} + \eta\frac{D_sS_n}{Dt} - \frac{\eta S_w}{\rho_s}\frac{D_s\rho_s}{Dt} = -\nabla\mathbf{q}_{rn} \tag{1.25}$$

**Assumption 1.5.1.** At the microscopic level the solid phase microscopic volume $dU$ remains constant, hence the solid phase can be considered incompressible.

$$\frac{Ddm_s}{Dt} := \frac{D(\rho_s\,dU_s)}{Dt} = \rho_s\overbrace{\frac{D(dU_s)}{Dt}}^{0} + dU_s\frac{D\rho_s}{Dt} = 0$$

From this, it is clear that $\dfrac{D\rho_s}{Dt} = 0$.

By considering assumption 1.5.1, equations (1.24) and (1.25) can be reduced to

$$\frac{\eta S_w}{\rho_w}\frac{D_w\rho_w}{Dt} + \frac{S_w}{1-\eta}\frac{D_s\eta}{Dt} + \eta\frac{D_sS_w}{Dt} = -\nabla\mathbf{q}_{rw} \tag{1.26}$$

for the wetting phase, and

$$\frac{\eta S_w}{\rho_n}\frac{D_n\rho_n}{Dt} + \frac{S_n}{1-\eta}\frac{D_s\eta}{Dt} + \eta\frac{D_sS_n}{Dt} = -\nabla\mathbf{q}_{rn} \tag{1.27}$$

for the non-wetting phase.

For practical reasons, it is convenient to write all material derivatives relative to the solid phase. Multiplying equation (1.26) and (1.27) by $\rho_w$, together with the definition of relative specific discharge

$$\eta S_w\frac{D_s\rho_w}{Dt} + \frac{S_w\rho_w}{1-\eta}\frac{D_s\eta}{Dt} + \eta\rho_w\frac{D_sS_w}{Dt} = -\nabla(\rho_w\mathbf{q}_{rw}) \tag{1.28}$$

$$\eta S_n\frac{D_s\rho_n}{Dt} + \frac{S_n\rho_n}{1-\eta}\frac{D_s\eta}{Dt} + \eta\rho_n\frac{D_sS_n}{Dt} = -\nabla(\rho_n\mathbf{q}_{rn}) \tag{1.29}$$

**Assumption 1.5.2.** The non-wetting phase has a constant and uniform pressure such that $p_n = 0$. Because equations (1.28) and (1.29) are a coupled system which is related by pressures $p_n$ and $p_w$. The system is decoupled and one of the equations can be neglected by this assumption.

**Assumption 1.5.3.** At the macroscopic level, the solid matrix is immobile, hence $\mathbf{V}_s = 0$. $\frac{\partial\eta}{\partial t} = 0$. Hence, equation (1.23) is reduced to

$$\frac{D_sG}{Dt} = \frac{\partial G}{\partial t} + \underbrace{\mathbf{V}_s\nabla G}_{=0} \Rightarrow \frac{D_sG}{Dt} = \frac{\partial G}{\partial t}$$

By assumption 1.5.2 the mass conservation equation for the non-wetting phase has been neglected and only the wetting-phase remains of interest. Finally, equation (1.28) is reduced to

$$\eta S_w\frac{\partial\rho_w}{\partial t} + \eta\rho_w\frac{\partial S_w}{\partial t} = -\nabla(\rho_w\mathbf{q}_{rw}) \tag{1.30}$$

which describes flow of a wetting-phase within a non-deformable porous medium in partially saturated conditions, where the wetting-phase is immiscible with the non-wetting phase which is assumed at constant pressure and no sinks/sources are considered.

Saturated conditions imply $S_w = 1$ and $\frac{\partial S_w}{\partial t} = 0$, which results in the equation for saturated flow in non-deformable porous media, with the same restrictions as for equation (1.30),

$$\eta\frac{\partial\rho_w}{\partial t} = -\nabla(\rho_w\mathbf{q}_{rw}) \tag{1.31}$$

## 1.6 Macroscopic momentum equation

Consider equation (1.16) for momentum, hence $E = m\mathbf{V}$ and $e = \rho\mathbf{V}$. By decomposing the total momentum flux in terms of the momentum flux relative to velocity: $\mathbf{j} = \rho\mathbf{V} + \mathbf{j}^M$. Note

that the momentum flux relative to mass velocity $\mathbf{j}^M$ is actually stress, hence $\mathbf{j}^M = -\boldsymbol{\sigma}$. Furthermore, momentum generation is $\Gamma^M = \mathbf{F}$ where $\mathbf{F}$ is the external body force acting on the phase.

$$\frac{\partial\left(\theta\overline{\rho\mathbf{V}}^{\alpha}\right)}{\partial t} = -\nabla\left[\theta\left(\overline{\rho\mathbf{V}}^{\alpha}\overline{\mathbf{V}}^{\alpha} + \overline{(\rho\mathring{\mathbf{V}})\mathring{\mathbf{V}}}^{\alpha}\right) - \overline{\boldsymbol{\sigma}}^{\alpha}\right) - \overbrace{[\rho\mathbf{V}\left(\mathbf{V}-\mathbf{u}\right)+\boldsymbol{\sigma})]\hat{\mathbf{n}}}^{\alpha\beta}\Sigma_{\alpha\beta} + \theta\overline{\rho\mathbf{F}}^{\alpha} \quad (1.32)$$

Using the identity (which is derived from the definitions of deviation and averages)

$$\overline{(\rho\mathring{\mathbf{V}})\mathring{\mathbf{V}}}^{\alpha} = \overline{\mathbf{V}}^{\alpha}\overline{\mathring{\rho}\mathring{\mathbf{V}}}^{\alpha} + \overline{\rho}^{\alpha}\overline{\mathring{\mathbf{V}}\mathring{\mathbf{V}}}^{\alpha} + \overline{\mathring{\rho}\mathring{\mathbf{V}}\mathring{\mathbf{V}}}^{\alpha}$$

on the right side of equation (1.32) and applying equation (1.9) to the left side, yields

$$\frac{\partial\left(\theta\overline{\rho}^{\alpha}\overline{\mathbf{V}}^{\alpha}\right)}{\partial t} + \frac{\partial\left(\theta\overline{\mathring{\rho}\mathring{\mathbf{V}}}^{\alpha}\right)}{\partial t} = -\nabla\left[\theta\left(\overline{\rho\mathbf{V}}^{\alpha}\overline{\mathbf{V}}^{\alpha} + \overline{\mathbf{V}}^{\alpha}\overline{\mathring{\rho}\mathring{\mathbf{V}}}^{\alpha} + \overline{\rho}^{\alpha}\overline{\mathring{\mathbf{V}}\mathring{\mathbf{V}}}^{\alpha} + \overline{\mathring{\rho}\mathring{\mathbf{V}}\mathring{\mathbf{V}}}^{\alpha} - \overline{\boldsymbol{\sigma}}^{\alpha}\right)\right] - \overbrace{[\rho\mathbf{V}\left(\mathbf{V}-\mathbf{u}\right)+\boldsymbol{\sigma})]\hat{\mathbf{n}}}^{\alpha\beta}\Sigma_{\alpha\beta} + \theta\overline{\rho\mathbf{F}}^{\alpha}$$

By combining with the mass conservation equation (1.18), manipulating somewhat and in indicial notation,

$$\theta\overline{\rho}^{\alpha}\frac{\partial\left(\overline{V_i}^{\alpha}\right)}{\partial t} + \frac{\partial\left(\theta\overline{\mathring{\rho}\mathring{V}_i}^{\alpha}\right)}{\partial t} = -\nabla\left[\theta\left(\overline{\mathring{\rho}\mathring{V}_i}^{\alpha}\overline{V_j}^{\alpha} + \overline{\rho}^{\alpha}\overline{\mathring{V}_i\mathring{V}_j}^{\alpha} + \overline{\mathring{\rho}\mathring{V}_i\mathring{V}_j}^{\alpha}\right)\right] + \nabla\left(\theta\overline{\sigma_{ij}}^{\alpha}\right) + \theta\overline{\rho F_i}^{\alpha} - \theta\frac{\partial\overline{V_i}^{\alpha}}{\partial x_j}\left(\overline{\rho}^{\alpha}\overline{V_j}^{\alpha} + \overline{\mathring{\rho}\mathring{V}_j}^{\alpha}\right) - \overbrace{\left[\rho\mathring{V}_i\left(V_i-u_j\right)-\sigma_{ij}\right]\hat{n}_j}^{\alpha\beta}\Sigma_{\alpha\beta}$$

**Assumption 1.6.1.** $\rho$ is constant

**Assumption 1.6.2.** $S_{\alpha\beta}$ is a material surface with respect to $\alpha$-phase

**Assumption 1.6.3.** Flow is macroscopically uniform: $\dfrac{\partial\overline{V_i}^{\alpha}}{\partial x_j} = 0$

Considering these assumptions, allows for

$$\theta\overline{\rho}^{\alpha}\frac{\partial\overline{V}^{\alpha}}{\partial t} = -\overline{\rho}^{\alpha}\nabla\left(\theta\overline{\mathring{\mathbf{V}}\mathring{\mathbf{V}}}^{\alpha}\right) + \nabla\left(\theta\overline{\boldsymbol{\sigma}}^{\alpha}\right) + \theta\overline{\rho\mathbf{F}}^{\alpha} + \overbrace{\boldsymbol{\sigma}\hat{\mathbf{n}}}^{\alpha\beta}\Sigma_{\alpha\beta}$$

**Assumption 1.6.4.** Dispersive mass fluxes are much smaller than advective mass fluxes $|\overline{\mathring{\rho}\mathring{\mathbf{V}}}^{\alpha}| \ll |\overline{\rho}^{\alpha}\overline{\mathbf{V}}^{\alpha}|$, thus may be considered as negligible.

**Assumption 1.6.5.** Dispersive momentum fluxes are much smaller than advective momentum fluxes $|\overline{(\rho\mathring{\mathbf{V}})\mathring{\mathbf{V}}}^{\alpha}| \ll |\overline{\rho\mathbf{V}}^{\alpha}\overline{\mathbf{V}}^{\alpha}| \approx |\overline{\rho}^{\alpha}\overline{\mathbf{V}}^{\alpha}\overline{\mathbf{V}}^{\alpha}|$.

With the aforementioned assumptions it is possible to write the macroscopic momentum balance equation:

$$\theta\overline{\rho_{\alpha}}^{\alpha}\frac{\partial\overline{\mathbf{V}_{\alpha}}^{\alpha}}{\partial t} + \theta\overline{\rho}^{\alpha}\overline{\mathbf{V}}^{\alpha}\nabla\overline{\mathbf{V}}^{\alpha} = \nabla\left(\theta\overline{\boldsymbol{\sigma}}^{\alpha}\right) + \overbrace{\boldsymbol{\sigma}\hat{\mathbf{n}}}^{\alpha\beta}\Sigma_{\alpha\beta} + \theta\,\overline{\rho_{\alpha}\mathbf{F}}^{\alpha}$$

which because of (1.15) is

$$\theta\overline{\rho_\alpha}^{\alpha}\frac{\partial\overline{V_i}^{\alpha}}{\partial t}+\theta\overline{\rho}^{\alpha}\overline{V_j}^{\alpha}\frac{\partial\overline{V_i}^{\alpha}}{\partial x_j}=\theta\overline{\frac{\partial\sigma_{ij}}{\partial x_j}}^{\alpha}+\theta\,\overline{\rho_\alpha F_i}^{\alpha} \tag{1.33}$$

At this point, it is necessary to clarify that, although variably saturated flow in porous media is in fact a multiphase system, the following reasoning is not one of a "true" multiphase system, since water flow will be thought of as uncoupled from air flow, in accordance to assumption 1.5.2. The analysis can be done as in fully coupled multiphase system, but it is unnecesary for the intended model, hence it may be approximated by a saturated analysis.

Assuming microscopical isochoric motion, evaluating the stress tensor, introducing the no-slip condition in the solid-fluid surface and rearranging [5],

$$\rho\Big[\underbrace{\frac{\partial q_i}{\partial t}}_{①}+\underbrace{\frac{\partial}{\partial x_j}\left(\frac{q_iq_j}{\eta}\right)}_{②}\Big]=-\eta\left(\frac{\partial p}{\partial x_j}+\rho g\frac{\partial x_3}{\partial x_i}\right)T^*_{ji}+\underbrace{\mu\frac{\partial^2 q_{ri}}{\partial x_j^2}}_{③}-\underbrace{-\mu\alpha_{ij}\frac{C_f}{\Delta_f^2}q^m_{rj}}_{④} \tag{1.34}$$

Where $p$ is fluid pressure, $q$ is specific discharge, $\eta$ is porosity, $\mu$ is viscosity, $T^*_{ij}$ and $\alpha_{ij}$ are tensorial properties fo the configuration of the solid-fluid surface when saturated with the phase of interest, $C_f$ is a macroscopic dimensionless shape factor and $\Delta_f$ is the ratio of void space volume to interface surface area. Term ① is the temporal variation, ② describes inertial forces, term ③ expresses the viscous forces due to shear inside the fluid and term ④ expresses the drag at the solid-fluid surfaces. Equation (1.34) can be transformed into a dimensionless form [5], from which the Reynolds (Re), Darcy (Da) and Strouhal (St) numbers for porous media can be formulated. The Reynolds number is a dimensionless ratio of inertial and viscous forces. Darcy's law is considered valid when $\mathrm{Re}\leq 1-10$ which is usually true in groundwater flows [6]. Darcy's number is a ratio of characteristic permeability, and the characteristic travel paths and distance. The Darcy number allows to estimate the magnitude of viscous resistance within the fluid, which is also related with the Reynolds number. The Strouhal number is the ratio between a characteristic travel time required to encounter a significant spatial change in velocity and a characteristic time required to encounter the same change in velocity (in time) in a mathematical point. In a sense, it can be interpreted as a ratio of local and convective accelerations.

Consider no momentum transfer between fluid phases and considering inertial effects and resistance of flow from viscous shear inside each fluid neglegible with respect to the shear produced with the solid. In equation (1.34) this means ② $\ll$ ④ and ③ $\ll$ ④. This conditions are true when $\mathrm{Re}\,\mathrm{Da}^{\frac{1}{2}}\ll 1$. Furthermore when the Strouhal number is small, $\mathrm{St}\leq 1$, term ① may be neglected too. Then equation 1.34 reduces to

$$q^m_{rj}=-\frac{1}{\mu}\underbrace{\frac{\Delta_f^2\eta}{C_f}\left(\alpha_{ji}\right)^{-1}T^*_{il}}_{k_{jl}}\left(\frac{\partial\overline{p}^f}{\partial x_j}+\overline{\rho}^f g\frac{\partial x_3}{\partial x_i}\right)$$

Assuming that tensors $(\alpha_{ij})^{-1}$ and $T^*_{il}$ have the same principal directions, and assuming cartesian coordinates, yields

$$\mathbf{q}_{r\alpha}=-\frac{\mathbf{k}_\alpha}{\mu_\alpha}\left(\nabla p_\alpha+\rho_\alpha g\nabla z\right) \tag{1.35}$$

where $\mathbf{k}_\alpha$ is the effective permeability to the $\alpha$-phase, and is a function of saturation. For the case of unsaturated water flow in soils, the air phase is usually considered with constant atmospheric pressure in the entire domain, hence an air phase equation in the manner of equation (1.35) is unnecesary. Only an equation for the wetting phase (i.e., water) is required.

**Assumption 1.6.6.** Water viscosity and density remains (approximately) constant.

For the water flow in soil, the effective hydraulic conductivity can be defined as

$$\mathbf{K}_w = \frac{\mathbf{k}_w g \rho_w}{\mu_w}$$

Piezometric head of water can be defined as $h_w = \dfrac{p_w}{g\rho_w}$.

By using such definitions, equation (1.35) for water specific discharge in a soil may be written as

$$\mathbf{q}_{rw} = -\mathbf{K}_w \nabla \left( h_w + z \right) \tag{1.36}$$

Note that hydraulic conductivity $\mathbf{K}$ derives from permeability $\mathbf{k}$ which was defined from a saturated, single fluid phase approach. In section 1.8 models for $\mathbf{K}$ will be considered which depend on such saturated approach and modified by factors associated to variable saturation.

## 1.7 Richards' Equation

Combining equation (1.30), which describes mass conservation in a non-deformable porous medium, and (1.36) which describes momentum conservation, leads to

$$\eta S_w \frac{\partial \rho_w}{\partial t} + \eta \rho_w \frac{\partial_s S_w}{\partial t} = \nabla \Big[ \rho_w \mathbf{K}_w \nabla \left( h_w + z \right) \Big]$$

Considering an incompressible wetting phase, such as water, results in Richards' equation [30] (phase subindices have been dropped for simplicity in notation),

$$\eta \frac{\partial S}{\partial t} = \nabla \Big[ \mathbf{K} \nabla \left( h + z \right) \Big]$$

Considering the definition of saturation, it is possible to write Richards' equation in terms of the wetting phase fraction $\theta$ $[L^3/L^3]$, which in common groundwater terminology is called volumetric water content.

$$\frac{\partial \theta}{\partial t} = \nabla \Big[ \mathbf{K} \nabla \Big( h + z \Big) \Big] \tag{1.37}$$

This equation relates the changes in water content with the primary driving forces: gravitational potential and pressure gradient and the properties of the porous medium, by means of its conductivity. From the mathematical point of view, Richards' equation is a parabolic equation in unsaturated regime, and an elliptic equation in saturated regime [18]. Richards' equation may be written in three forms depending on the choice of variables. These are shown here in 1D form. They are the *water content* form in terms of water content $\theta$, conductivity $K(\theta)$ and diffusivity $D(\theta)$

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z} \left[ D(\theta) \frac{\partial \theta}{\partial z} \right] + \frac{\partial K(\theta)}{\partial z} \tag{1.38}$$

the *pressure* or *matric* form in terms of pressure $h$, hydraulic capacity $C(h)$ and conductivity $K(h)$

$$C(h)\frac{\partial h}{\partial t} = \frac{\partial}{\partial z}\left[K(h)\left(\frac{\partial h}{\partial z} + 1\right)\right] \tag{1.39}$$

and the *mixed* form in terms of pressure $h$, water content $\theta$ and conductivity $K(h)$.

$$\frac{\partial \theta(h)}{\partial t} = \frac{\partial}{\partial z}\left[K(h)\left(\frac{\partial h}{\partial z} + 1\right)\right] \tag{1.40}$$

where $C = \dfrac{\partial \theta}{\partial h}$ $[1/L]$ is the hydraulic capacity of the soil and $D = \dfrac{K}{C}$ $[L^2/T]$ the diffusivity. This definitions allow to relate all three forms.

Although it is implicit in the equations, it is worth to observe that specific discharge $\mathbf{q}_w$ can be thought of as the darcian velocity or flux $J$, which is positive upwards because of the adopted sign convention.

The three forms (1.38) (1.39) (1.40) have different properties. The water content form is the conservative form, in the sense that the variable of interest is a conserved variable. Because of this it shows very good conservation properties [13]. However, water content varies only when in the unsaturated region, hence the equation is useless when flow occurs in a saturated regime. From a conceptual perspective, these equations do not show explicitly the driving forces of flow, since it is the pressure gradient which generates flow, which is associated to different water contents by the soil constitutive model.

The pressure form involves only changes in pressure. Although, when coupled with the soil constitutive model, it also relates to water content. Because pressure is a continuous function from negative pressure (suction, matric potential) in an unsaturated regime to positive pressures in a saturated regime, the transiton is well handled by solving pressure. On the other hand, this equation is not written in terms of a conserved variable and, when solved by numerical methods, may show problems in conservation [13].

The mixed form relates the change in water content to the pressure gradients. It is, in the conceptual sense, better to understand the driving forces of flow. Furthermore, because the conserved variable is present in the equation, conservation is better handled with this equation [13] [29]. As written in equation (1.40) it is not continuous into the saturated region, since water content becomes constant.

Note that from the physical perspective, all three forms model the same phenomenon. Mathematically, from the differential point of view conservation is not an issue, but from the numerical perspective it is, as will be discussed in the following chapter. Because in this work variably saturated soils are of primary interest the water content form is not used. It is also important to observe that all forms depend strongly on $K(h)$ or $K(\theta)$, $C(h)$, $D(h)$ or $D(\theta)$ functions defined by a constitutive model for the porous media, which is essential.

## 1.8 Unsaturated soil constitutive model

Richards' equation in any form requires the hydraulic conductivity $K(h)$ function and the water content $\theta(h)$ function to be known. These functions interrelate pressure, water content, conductivity and other soil properties. There are several models [34] that feed upon soil parameters to generate mathematical relations for the functions $K(h)$, $\theta(h)$ and its derivative,

$C(h)$. One of the best known and often used is the Mualem-van Genuchten model [37]. Other well known models are those of Brooks-Corey [9] and Gardner-Russo [16] [32].

### 1.8.1 Mualem-van Genuchten Model

The Mualem-van Genuchten model (MG) is expressed by

$$\theta = \begin{cases} \dfrac{\theta_s - \theta_r}{\left(1 + (\alpha|h|)^{\hat{\eta}}\right)^{\mu}} + \theta_r & \text{if } h \leq 0 \\ \theta_s & \text{if } h > 0 \end{cases} \tag{1.41}$$

$$K = \begin{cases} \dfrac{\left[1 - (\alpha|h|)^{\hat{\eta}-1}\left(1 + (\alpha|h|)^{\hat{\eta}}\right)^{-\mu}\right]^2}{\left(1 + (\alpha|h|)^{\hat{\eta}}\right)^{\mu/2}} K_s & \text{if } h \leq 0 \\ K_s & \text{if } h > 0 \end{cases} \tag{1.42}$$

$$C = \begin{cases} \dfrac{\partial\theta}{\partial h} = -\mu\hat{\eta}\alpha(\theta_s - \theta_r)(1 + \alpha^{\hat{\eta}}|h|^{\hat{\eta}})^{-\mu-1}h & \text{if } h \leq 0 \\ 0 & \text{if } h > 0 \end{cases} \tag{1.43}$$

$$\mu = 1 - \frac{1}{\hat{\eta}} \tag{1.44}$$

Where $\theta_r\,[L^3/L^3]$ is the residual water content, $K_s\,[L/T]$ is the saturated hydraulic conductivity, $\hat{\eta}$ is a parameter which measures pore-size distribution and $\alpha\,[L^{-1}]$ is a parameter related to the inverse of the air-entry pressure. Note that the hydraulic capacity function $C(h)$ is obtained by analytical differentiation from the water content function. It is possible to approximate $C$ in a discrete way, but it has been shown [29] that the analytical approach is accurate and efficient. It is important to note that water content and conductivity have a maximum value at saturation, and that the hydraulic capacity is zero at saturation and has a maximum value for a certain suction pressure which can be seen in figure 1.3(a) and with further detail near saturation in 1.3(b).

### 1.8.2 Modified Mualem-van Genuchten Model

Because of the non-linear behavior of the hydraulic conductivity function $K(h)$, in particular in the range close to saturation errors that can be attributed to the approximation of the conductivity arise [39]. This is especially important when $\hat{\eta} < 2$, which corresponds to fine soils or undisturbed soils with broad pore-size distributions. Vogel et al. [39] suggest that a parameter $h_s$ should be included in the formulation of the MG model in order to better describe effects that seem to be important in fine soils. This parameter $h_s$, although artificial, somehow simulates air-entry, and minimum capillary height for macropores [39]. The practical effect is that the highly non-linear behavior of $K$ near saturation is turned into a constant value in a smooth fashion, which can be seen in figure 1.3(b) (for $h_s = 4\,cm$). This has further benefits such as better stability around saturation [33] [39]. Although Vogel et al. suggest that $h_s \approx -2cm$, further studies by Schaap and Van Genuchten [33] show that $h_s = -4cm$ better represents $K(h)$ near saturation. Børgesen et al. proposed a different modification using a scaling function [12] with a parameter $h_m$ which they interpreted as the boundary between macropore flow and matrix flow, and found their parameter to be optimal

around $h_m = -4cm$. The proposed Modified Mualem-van Genuchten model (MMG) [39] [33] is shown in the following equations.

$$\theta_m = \begin{cases} \Big(\theta_s - \theta_r\Big)\Big(1 + (\alpha|h_s|)^{\hat{\eta}}\Big)^{\mu} + \theta_r & \text{if } h \leq h_s \\ \theta_s & \text{if } h > h_s \end{cases} \tag{1.45}$$

$$S_e = \frac{\theta - \theta_r}{\theta_s - \theta_r} \tag{1.46}$$

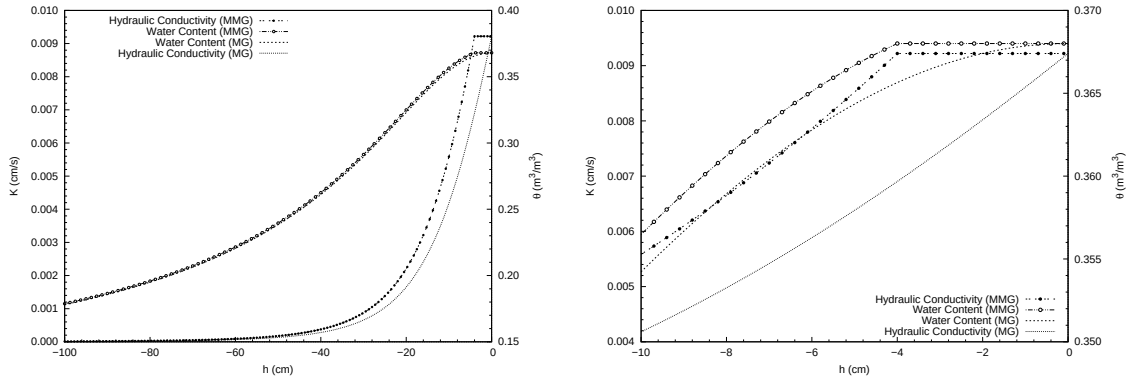
$$\theta = \frac{\theta_m - \theta_r}{\Big(1 + (\alpha|h|)^{\hat{\eta}}\Big)^{\mu}} + \theta_r \tag{1.47}$$

$$K = \begin{cases} K_s S_e^{1/2} \dfrac{1 - \left[1 - \left(\frac{\theta_s - \theta_r}{\theta_m - \theta_r} S_e\right)^{\frac{1}{\mu}}\right]^{\mu}}{1 - \left[1 - \left(\frac{\theta_s - \theta_r}{\theta_m - \theta_r}\right)^{\frac{1}{\mu}}\right]^{\mu}} & \text{if } h \leq h_s \\ K_s & \text{if } h > h_s \end{cases} \tag{1.48}$$

$$C = \begin{cases} \dfrac{\partial \theta}{\partial h} = -\mu\hat{\eta}\alpha(\theta_m - \theta_r)(1 + \alpha^{\hat{\eta}}|h|^{\hat{\eta}})^{-\mu-1}h & \text{if } h \leq h_s \\ 0 & \text{if } h > h_s \end{cases} \tag{1.49}$$

$$\mu = 1 - \frac{1}{\hat{\eta}} \tag{1.50}$$

Note that if $h_s = 0$ the MMG model is reduced to the MG model. Figure 1.3 compares the MG and the MMG models for a particular soil, with $h = -4\ cm$. The figure shows that although the $K(h)$ function is evidently different, the $\theta(h)$ function is very similar in both models, hence mass conservation is unaffected.




(a) (b)

Figure 1.3: Soil properties, MG and MMG models with $h_s = 4\,cm$

# Chapter 2

# Numerical model

Richards' equation has no general analytical solution, hence, the use of numerical approximations is necessary. In this chapter, several numerical schemes are presented, formulated for 1D vertical flow, which allows to study the numerical consequences of each choice without further complexity and computational cost generated by multidimensional problems.

Approximations for the solution of the mixed form (1.40) and the pressure form (1.39) are formulated. Schemes have been developed for the pressure and mixed forms of Richards' equation only since the water-content form is of no use when simulating saturated conditions. Depending on the case either $h$ or $\theta$ may be solved by the scheme, and the constitutive model allows for solution of the other variable.

In the following sections the schemes are presented and discussed in terms of the equation they approximate, their scope and range of applicabilty, as well as stability and other properties. Scheme formulations are presented in Appendix A and details on stability analysis are presented in Appendix B.

## 2.1 Spatial and temporal discretization

The spatial derivatives have been approximated with a centered finite difference scheme summarized in figure 2.1. Subscript $i$ is the spatial index such that $1 \leq i \leq N$, where $i = 1$ is the lower boundary cell in the discrete porous stratum and $i = N$ is the uppermost boundary cell. Hydraulic conductivity is evaluated at the boundaries between cells $K_{i\pm1/2}$, not the cells themselves as it is a flux coefficient, and fluxes are estimated at the cell interfaces. Because $K$ is function of $h$ or $\theta$ it can be computed in the cells themselves. This raises the issue of intercell conductivity $K_{i\pm1/2}$ estimation.

For the time derivative both explicit and implicit schemes are developed, by means of first order forward and backward Euler methods.

## 2.2 Explicit formulation

The simplest way to discretize the time derivative in equation (1.40) is by a forward Euler scheme, which yields an explicit formulation for the solution. The advantages of explicit schemes include that they are simple and straightforward and that mass conservation is, when

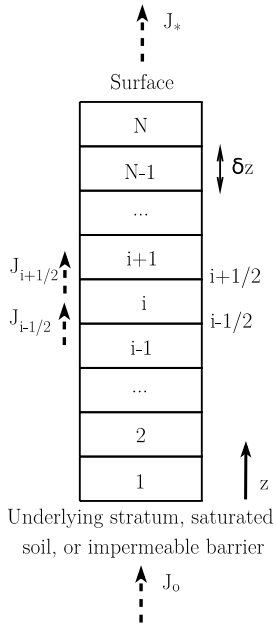



Figure 2.1: Discrete domain

formulated correctly, excellent. However, such schemes have conditional stability subject to time step limits. From the mixed form (1.40) it is possible to formulate

$$\frac{\theta^{n+1}-\theta^{n}}{\Delta t}-\frac{\partial}{\partial z}\left[K(h)\frac{\partial h}{\partial z}+K(h)\right]^{n}=0 \tag{2.1}$$

and from the pressure form (1.39)

$$C^{n}\frac{h^{n+1}-h^{n}}{\Delta t}-\frac{\partial}{\partial z}\left[K(h)\frac{\partial h}{\partial z}+K(h)\right]^{n}=0 \tag{2.2}$$

When evaluated accordingly, the EMC and EP schemes are formulated.

### 2.2.1 Mixed scheme (EMC)

From equation (2.1), the *Explicit Mixed Conservative* (EMC) scheme is as follows,

$$\theta_i^{n+1}=\frac{\Delta t}{\delta z^2}K_{i+1/2}^{n}h_{i+1}^{n}-\frac{\Delta t}{\delta z^2}\left(K_{i+1/2}^{n}+K_{i-1/2}^{n}\right)h_i^{n}+\frac{\Delta t}{\delta z^2}K_{i-1/2}^{n}h_{i-1}^{n}+\frac{\Delta t}{\delta z}\left(K_{i+1/2}^{n}-K_{i-1/2}^{n}\right)+\theta_i^{n} \tag{2.3}$$

### 2.2.2 Pressure-based scheme (EP)

From equation (2.2), the *Explicit Pressure-based* (EP) scheme is as follows,

$$h_i^{n+1}=\frac{\Delta t}{C_i^{n}\delta z^2}K_{i+1/2}^{n}h_{i+1}^{n}-\frac{\Delta t}{C_i^{n}\delta z^2}\left(K_{i+1/2}^{n}+K_{i-1/2}^{n}\right)h_i^{n}+\frac{\Delta t}{C_i^{n}\delta z^2}K_{i-1/2}^{n}h_{i-1}^{n}+\frac{\Delta t}{C_i^{n}\delta z}\left(K_{i+1/2}^{n}-K_{i-1/2}^{n}\right)+h_i^{n} \tag{2.4}$$

### 2.2.3 Boundary conditions

For Dirichlet conditions, only $h_1^{n+1}$ and/or $h_N^{n+1}$ need to be imposed. No further treatment is required because $C$ and $K$ are evaluated in time $n$. For Neumann conditions the terms that need to be supplied are those related to the vertical pressure gradient $\frac{\partial h}{\partial z}$. For a known flux $J^*$ at the upper boundary, the expression for $h_N^{n+1}$ is

$$h_N^{n+1} = h_N^n - \frac{\Delta t}{C_N^n \delta z}\left[J^* + K_{N-1/2}^n\left(\frac{h_N^n - h_{N-1}^n}{\delta z} + 1\right)\right] \tag{2.5}$$

Conversely, for a known flux $J_o$ at the lower boundary, the expression for $h_1^{n+1}$ is

$$h_1^{n+1} = h_1^n + \frac{\Delta t}{C_1^n \delta z}\left[J_o + K_{1+1/2}^n\left(\frac{h_2^n - h_1^n}{\delta z} + 1\right)\right] \tag{2.6}$$

Clearly, an impervious stratum can be simulated by assigning $J^* = 0$ and/or $J_o = 0$ as needed. Gravity flow in a semi-infinite stratum can be simulated simply by imposing $h_1^{n+1} = h_2^{n+1}$, which ensures a null pressure gradient.

## 2.3 Implicit formulation

From the pressure form (1.39) it is possible to write an implicit discretization technique

$$C(h^{n+1})\frac{h^{n+1} - h^n}{\Delta t} = -\frac{\partial}{\partial z}\left[K(h\left(\frac{\partial h}{\partial z}\right)\right]^{n+1} \tag{2.7}$$

and from the mixed form (1.40),

$$\frac{\theta^{n+1} - \theta^n}{\Delta t} = -\frac{\partial}{\partial z}\left[K(h\left(\frac{\partial h}{\partial z}\right)\right]^{n+1} \tag{2.8}$$

which give rise to the two implicit schemes that are presented in this section.

Because of the implicit formulation and the high non-linearity of $K(h)$ and in the case of the IMC scheme also $\theta(h)$, some method of linearization is required. Picard and Newton iteration schemes are good choices. Several authors [15] [25] [26] have concluded that the Newton scheme may be more efficient in some cases than the Picard scheme and even converge when the Picard scheme does not, but in other cases it may converge to the wrong solution. Hence, the Picard scheme is preferred in this work. In terms of notation, iterations are denoted by supercript $m$. In both cases, $K$ and $C$ are approximated by Picard iteration in time $(n+1, m)$, when solving for pressure in time $(n+1, m+1)$.

### 2.3.1 IP Scheme

A simple implicit scheme can be written from equation (2.7), the Implicit Pressure based (IP) Scheme,

$$a_i h_{i-1}^{n+1,m+1} + b_i h_i^{n+1,m+1} + c_i h_{i+1}^{n+1,m+1} = f_i \tag{2.9}$$

where the coefficients are

$$a_i = -\frac{\Delta t}{\delta z^2} K_{i-1/2}^{n+1,m} \tag{2.10}$$

$$b_i = C_i^{n+1,m} + \frac{\Delta t}{\delta z^2}\left(K_{i-1/2}^{n+1,m} + K_{i+1/2}^{n+1,m}\right) \tag{2.11}$$

$$c_i = -\frac{\Delta t}{\delta z^2} K_{i+1/2}^{n+1,m} \tag{2.12}$$

$$f_i = C_i^{n+1,m} h_i^n + \frac{\Delta t}{\delta z}\left(K_{i+1/2}^{n+1,m} - K_{i-1/2}^{n+1,m}\right) \tag{2.13}$$

### 2.3.2 IMC scheme

From the discretized mixed-form (2.8), but actually solving for pressure the Implicit Mixed Conservative (IMC) scheme is formulated. Solving for pressure allows for the scheme to transition from unsaturated to saturated regimes, hence it is actually a solution for *variably* saturated flows. The highly non-linear nature of $C$ is a key factor in mass conservation when solving pressure, which is not a conserved variable. Although it is correct to define $C = \frac{\partial \theta}{\partial h}$ when dealing with continous functions, it is necessary to be carefull when evaluating them in a discrete way. The approximation

$$\frac{\partial \theta}{\partial t} \approx \frac{\theta^{n+1,m+1} - \theta^n}{\Delta t} \approx C^{n+1,m} \frac{h^{n+1,m+1} - h^n}{\Delta t}$$

disregards the fact that $C$ is a function of $h$ and hence a function of $t$, which in the discrete form is not an accurate approximation. The time derivative of $\theta$ should consider the chain rule for the time derivative of $h$, both in time stepping form $n$ to $n+1$ and from $m$ to $m+1$. If this derivative is not considered the scheme shows poor mass conservation, as clearly shown by Celia et al. [13]. In order to consider this, Celia et al. showed that the use of the first order Taylor polynomial applied to the time derivative of $\theta$, around $h^{n+1,m}$ is a good solution, for the method becomes perfectly mass conservative [13].

$$\theta^{n+1,m+1} = \theta^{n+1,m} + C^{n+1,m}\left(h^{n+1,m+1} - h^{n+1,m}\right) \tag{2.14}$$

This results in a better approximation

$$\frac{\partial \theta}{\partial t} \approx \frac{\theta^{n+1,m+1} - \theta^n}{\Delta t} \approx \frac{\theta^{n+1,m} + C^{n+1,m}\left(h^{n+1,m+1} - h^{n+1,m}\right) - \theta^n}{\Delta t}$$

where the derivatives in the $m$ iteration are properly considered.

Writing the equation considering the Taylor polynomial and the Picard iteration results in the Implicit Mixed Conservative (IMC) scheme,

$$a_i h_{i-1}^{n+1,m+1} + b_i h_i^{n+1,m+1} + c_i h_{i+1}^{n+1,m+1} = f_i \tag{2.15}$$

Where coefficients $a, b, c$ and the term $f$ are

$$a_i = -\frac{\Delta t}{\delta z^2} K_{i-1/2}^{n+1,m} \tag{2.16}$$

$$b_i = C_i^{n+1,m} + \frac{\Delta t}{\delta z^2}\left(K_{i-1/2}^{n+1,m} + K_{i+1/2}^{n+1,m}\right) \tag{2.17}$$

$$c_i = -\frac{\Delta t}{\delta z^2} K_{i+1/2}^{n+1,m} \tag{2.18}$$

$$f_i = C_i^{n+1,m} h_i^{n+1,m} + \frac{\Delta t}{\delta z}\left(K_{i+1/2}^{n+1,m} - K_{i-1/2}^{n+1,m}\right) + \theta_i^n - \theta_i^{n+1,m} \tag{2.19}$$

This scheme was first proposed by Celia et al. [13] both for finite difference and finite element schemes. It is well-known and frequently used.

### 2.3.3 Boundary conditions

Dirichlet boundary conditions are treated as imposed pressure head conditions, while Neumman conditions are pressure gradients. Physically, imposing positive pressure head in the upper boundary can represent surface water height. Negative upper pressure head seems less natural. Imposed pressure head at the lower boundary is somewhat difficult to imagine, as it appears artificial to have pressure below the soil column which does not depend on the soil column. Neumman conditions allow to simulate a semi-infinite stratum in the soil, which responds only to the state of the column above it.

In order to impose flow or zero-flow (impervious) conditions, Richards' equation can be written in terms of flux, which allows to write the derivative of the flux, and evaluate fluxes in the intercell boundaries and to impose one of them while expressing the other in terms of the pressure gradient and conductivity as in the schemes. Flow conditions have a very clear physical meaning, even when the imposed flow is zero, which can represent impervious strata, or no infiltration from the upper boundary.

#### Imposed pressure head

When imposing pressure head at the upper and lower boundary conditions it is necessary to eliminate the first (in the case of $i = 1$, the lower boundary ) or last ($i = N$, the upper boundary) from the system of equations or matrix system, since $h_i^{n+1}$ or $h_N^{n+1}$ respectively, would not be unknowns, but would be part of term $f$. Then, it is only necessary to impose $h_i^{n+1} = h_{low}(t)$ or $h_N^{n+1} = h_{upp}(t)$.

Lower boundary

- IP

$$f_2 = C_2^{n+1,m} h_2^n + \frac{\Delta t}{\delta z}\left(K_{2+1/2}^{n+1,m} - K_{2-1/2}^{n+1,m}\right) - a_2 h_1^{n+1} \tag{2.20}$$

- IMC

$$f_2 = C_2^{n+1,m} h_2^{n+1,m} + \frac{\Delta t}{\delta z}\left(K_{2+1/2}^{n+1,m} - K_{2-1/2}^{n+1,m}\right) + \theta_2^n - \theta_2^{n+1,m} - a_2 h_1^{n+1} \tag{2.21}$$

Upper boundary

- IP

$$f_{N-1} = C_{N-1}^{n+1,m} h_{N-1}^{n} + \frac{\Delta t}{\delta z}\left(K_{(N-1)+1/2}^{n+1,m} - K_{(N-1)-1/2}^{n+1,m}\right) - c_{N-2} h_N^{n+1} \tag{2.22}$$

- IMC

$$f_{N-1} = C_{N-1}^{n+1,m} h_{N-1}^{n+1,m} + \frac{\Delta t}{\delta z}\left(K_{(N-1)+1/2}^{n+1,m} - K_{(N-1)-1/2}^{n+1,m}\right) + \theta_{N-2}^{n} - \theta_{N-2}^{n+1,m} - c_{N-2} h_N^{n+1} \tag{2.23}$$

**Imposed flow**

Flow at the boundaries is imposed at the lower face of cell $i = 1$ or the upper face of cell $i = N$. The first is denoted $J_o$ and the latter $J^*$, both positive upwards as shown in figure 2.1. The corresponding coefficients for the first or last row of the matrix equation are as shown below. Note that there is no $a_1$ term as there is no $h_0^{n+1}$ unknown at the lower boundary, and at the upper boundary there is no $c_N$ term as there is no $h_{N+1}^{n+1}$ unknown.

Lower boundary

$$b_1 = C_1^{n+1,m} + \frac{\Delta t}{\delta z^2} K_{1+1/2}^{n+1,m} \tag{2.24}$$

$$c_1 = -\frac{\Delta t}{\delta z^2} K_{1+1/2}^{n+1,m} \tag{2.25}$$

- IP

$$f_1 = C_1^{n+1,m} h_1^{n} + \frac{\Delta t}{\delta z}\left(K_{1+1/2}^{n+1,m}\right) + \frac{\Delta t}{\delta z} J_o^{n+1} \tag{2.26}$$

- IMC

$$f_1 = C_1^{n+1,m} h_1^{n+1,m} + \frac{\Delta t}{\delta z}\left(K_{1+1/2}^{n+1,m}\right) + \theta_1^{n} - \theta_1^{n+1,m} + \frac{\Delta t}{\delta z} J_o^{n+1} \tag{2.27}$$

Upper boundary

$$a_N = -\frac{\Delta t}{\delta z^2} K_{N-1/2}^{n+1,m} \tag{2.28}$$

$$b_N = C_N^{n+1,m} + \frac{\Delta t}{\delta z^2} K_{N-1/2}^{n+1,m} \tag{2.29}$$

- IP

$$f_N = C_N^{n+1,m} h_N^{n} - \frac{\Delta t}{\delta z}\left(K_{N-1/2}^{n+1,m}\right) - \frac{\Delta t}{\delta z} J_*^{n+1} \tag{2.30}$$

- IMC

$$f_N = C_N^{n+1,m} h_N^{n+1,m} - \frac{\Delta t}{\delta z}\left(K_{N-1/2}^{n+1,m}\right) + \theta_N^{n} - \theta_N^{n+1,m} - \frac{\Delta t}{\delta z} J_*^{n+1} \tag{2.31}$$

**Lower boundary: Gravity flow/Semi-infinite stratum**

In order to allow for free gravity flow, the pressure gradient must be zero at the boundary, and because of this, the hydraulic conductivity gradient is also zero. This means $h_1$=$h_2$, hence $K_1 = K_2$ at all times. Finally, flow will be equal to hydraulic conductivity. This can be written in terms of equation (2.37), for $i = 2$. However, in this particular case it is possible to rewrite coefficients $a$ and $b$ because $h_1 = h_2$, so that $h_1$ is not actually solved independently (hence $a_2^* = 0$, to remove it from the matrix solution), but later assigned as equal to the solution of $h_2$.

$$a_2^* h_1^{n+1,m+1} + b_2^* h_2^{n+1,m+1} + c_2 h_3^{n+1,m+1} = f_2 \tag{2.32}$$

$$a_2^* = 0 \tag{2.33}$$

$$b_2^* = a_2 + b_2 \tag{2.34}$$

**Impervious stratum**

In the case that a boundary is an impervious stratum or barrier, the flux at such boundary $J_{1/2}$ or $J_{N+1/2}$ must be imposed as zero. The corresponding coefficients for the first row of the matrix equation are the same as those of the known flow case, with $J_o(t) = 0$ for the lower boundary or $J_*(t) = 0$ for the upper boundary.

### 2.3.4 Convergence and under-relaxation

Because Picard iterations are performed in each time step to approximate $C$ and $K$, an appropriate convergence criterion is required. The standard is to stop iterations when convergence error $h_i^{n+1,m+1} - h_i^n$ is less than a specified convergence tolerance $\epsilon$.

Huang et al. [22] showed that the standard criterion, although effective, is not particularly efficient in terms of computational time. Huang et al. proposed a $\theta$-based criterion which is computationally more efficient, i.e. $\theta_i^{n+1,m+1} - \theta_i^{n+1,m}$. This type of criterion has been also used succesfully by Vanderborght et al. [38], although they found that the $\theta$-based criterion may lead to innacurate matric potential profiles if the value of the $C(h)$ function is small, in which case the standard criterion can be used. Similar experiences are reported by van Dam and Feddes [36]. Both criteria were implemented in this model, although for validation and comparison purposes the h-based criterion has been favored because of the aforementioned observations that the $\theta$-based method may generate errors near saturation.

Phoon et al. [28] studied the effects of under-relaxation to accelerate convergence, and conducted a comparison of two under-relaxation methods: UR1,

$$K_i^{n+1,m} = K\left(\frac{h_i^{n+1,m} + h_i^n}{2}\right) \tag{2.35}$$

and UR2

$$K_i^{n+1,m} = K\left(\frac{h_i^{n+1,m} + h_i^{n,m-1}}{2}\right) \tag{2.36}$$

Phoon et al. found that, although UR1 is faster than UR2 (and also faster than UR0, i.e, when no under-relaxation is used), it can generate inaccurate results. UR2 improves

convergence rates compared to no under-relaxation, to a lower degree than UR1 but is much more accurate. In the cases reported in this work, no significant differences in CPU time were found with either methods.

## 2.4 Scheme properties

### 2.4.1 Solution method

The explicit nature of the EMC and EP schemes imply that no iterations are necessary and each cell can be solved independently in time $n+1$, without the need of a matrix equation, despite the three point finite difference stencil and can be solved by directly evaluating terms in time $n$.

The implicit methods however require the solution of a matrix equation. It is possible to write the schemes (as has been noted already in their description) in a coefficient fashion:

$$a_i h_{i-1}^{n+1,m+1} + b_i h_i^{n+1,m+1} + c_i h_{i+1}^{n+1,m+1} = f_i \tag{2.37}$$

It is important to note that term $f$ includes the water content at time $n$, and all other terms correspond to time $n+1$, at the $m$ iteration. The unknown vector is the matric potential in the entire domain in time $n+1$ and $m+1$ iteration. By writing equation (2.37) from $i=1$ to $i=N$, a linear system can be obtained, with $N$ equations that can be written in matrix form

$$\begin{bmatrix} b_1 & c_1 & 0 & \cdots & 0 \\ a_2 & b_2 & c_2 & 0 & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & 0 & a_{N-1} & b_{N-1} & c_{N-1} \\ 0 & \cdots & 0 & a_N & b_N \end{bmatrix} \begin{bmatrix} h_1^{n+1,m+1} \\ h_2^{n+1,m+1} \\ \vdots \\ h_{N-1}^{n+1,m+1} \\ h_N^{n+1,m+1} \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \end{bmatrix} \tag{2.38}$$

This system is clearly tridiagonal and can be solved by Thomas Algorithm, which is an efficient solution of the system. Nevertheless, the implicit schemes require iteration. Hence, for a given time step, the Thomas Algorithm will need to be perfomed as many times as the iterative process requires to converge. In other words, every iteration requires solution of the entire system.

### 2.4.2 Transition from unsaturated to saturated

In neither the EMC nor EP formulations can continuity into the saturated region be achieved. These schemes can solve for the first unsaturated cell $i$ which becomes saturated in $n+1$. However, for the next time step the time derivative vanishes in the EMC scheme which in turn vanishes the unknown $\theta_i^{n+1}$ which was to be solved for. In the EP scheme $C_i^n \to 0$ (as the soil becomes saturated) and the equations are undefined. Hence, calculations must be halted whenever such conditions arise in a time step.

Conceptually, the inability of the EMC scheme to solve in saturated conditions is due to the fact that the water-content function is piece-wise defined. It is a smooth function for $h \leq 0$,

but it is a constant with value $\theta_s$ for $h \geq 0$. Hence, there is no difference in water content for an infinite range of positive pressures. The inability of the EP scheme is of the same nature, although it is through $C$. In this case, because the state variable to solve for is pressure, from such perspective there is no limitation to variably saturated solutions. But, because the water content function is constant for saturated conditions, then $C = 0$ in such conditions which undefines the equations. It is in fact the piece-wise definition of $\theta$ which leads to this.

Even so, for a one-dimensional case, the solution of the entire profile could be achieved even when saturation conditions arise. Because the suction profile can be determined by the model, the position of the water table can be obtained. Pressure distribution below the water table corresponds to a hydrostatic distribution, thus, the entire pressure profile can be described. Nevertheless, this poses a restriction for the domain of the problem which is indeed solved by the scheme, i.e., the unsaturated region, forcing to define $h \in ]-\infty, 0[$. This demands that the spatial domain changes size as to coincide with the previous restriction. This is only a computational nuisance, which requires appropriate treatment when coded.

The IP and IMC schemes can solve the entire domain even in variably saturated conditions. The issues that impede this in the EMC and EP schemes are not present in the implicit schemes, firstly, because the state variable to solve for is pressure, not water content. Pressure is a continuous function: $h \in ]-\infty, \infty[$ while the water content function is not. Still, water capacity $C$ is zero whenever saturation occurs. However, because of the implicit approach, $C$ is no longer a denominator, but a summand in the non-zero coefficient $b_i$ (main diagonal of the coefficient matrix) and in the constant term $f_i$.

In summary, to achieve variably saturated solutions, it is necessary to solve for pressure, either directly by discretizing the pressure form (as in the IP scheme) or the mixed form (as in the IMC scheme). However, it is not sufficient to solve for pressure, as can be seen with the EP scheme. To achieve continuity from unsaturated to saturated regimes, the implicit schemes are necessary.

### 2.4.3 Mass conservation

The explicit EMC scheme is a conservative scheme. It solves directly for the conserved state variable. The EP scheme, however, shows very poor conservation properties, despite de fact that it is an explicit scheme. The reason for this is that it solves for pressure head which is not a conserved variable, and relates it to mass conservation through the non-linear function $C$, which is poorly discretized in time in this scheme, given that the time derivative is evaluated without considering the non-linear relations between $C$, $\theta$ and $h$.

Comparison between equations (2.19) and (2.13) shows that only the $\theta$ terms are present in one and not in the other, and that in equation (2.19) the constant term is dependent on $h_i^{n+1,m}$ while in equation (2.13) it is dependent on $h_i^n$. Note that the difference between the IMC scheme and the IP scheme occurs in the constant term $f_i$. Furthermore, the formulation of boundary conditions changes exactly in the same manner, by applying the same variations in $f_i$. These are the terms responsible for adequate mass conservation [13], whilst all other terms remain identical.

The issues that affect the IP scheme which are solved by the IMC scheme, are of the same nature of those responsible for poor conservation in the EP scheme, and are related to the treatment of $C$ and the time derivative of $\theta(h)$. The pressure form solves for a non-conservative state variable, while the water content form, and the mixed, form solve for a conserved variable. From the work of [13] and onwards, the use of the mixed form as in the IMC scheme

has been widespread. Nevertheless, other approaches have been taken. Rathfelder and Abriola [29] developed mass conservative methods with the h-form by discretizing the hydraulic capacity function with standard chord slope aproximations, but also found that the mixed form, together with an analytical expression for the hydraulic capacity function was computationally more efficient. Phoon et al. [28] also found that under certain under-relaxation techniques, the h-form can be solved with good mass conservation, even in coarse grids.

Several authors [13] [22] [28] also emphasize the fact that mass conservation is a necessary condition for an accurate solution, but does not guarantee it. Other factors have severe influence over the accuracy, especially the shape of the hydraulic conductivity function and the effects of discretization on this function.

### 2.4.4 Stability

Stability analysis was performed for the EMC and the IMC schemes only, since basic properties such as mass conservation and accuracy are not well achieved with the EP or the IP scheme. Von Neumann analysis of small perturbations is used. In the case of the EMC scheme, water content perturbations $\tilde{\theta}$ of amplitude $b$ around a base value $a$ are done by

$$\tilde{\theta} = a + be^{\mathbf{i}\psi z} \tag{2.39}$$

In the case of the IMC scheme, pressure perturbations $\tilde{h}$ are studied.

$$\tilde{h} = a + be^{\mathbf{i}(\psi z - \omega t)} \tag{2.40}$$

Furthermore, because the equation is highly non linear, further assumptions are necessary:

**Assumption 2.4.1.** Water content is a linear function of pressure

$$\theta(h) = \theta_o + C(h - h_o) \tag{2.41}$$

**Assumption 2.4.2.** Hydraulic conductivity is a linear function of water content (EMC scheme) or pressure (IMC scheme).

$$K(\theta) = K_o + K_*(\theta - \theta_o) \tag{2.42}$$

$$K(h) = K_o + K_* h(h - h_o) \tag{2.43}$$

**Assumption 2.4.3.** The smallest discretized wave which can be observed in a mesh of size $\delta z$ is of wavelength $\lambda = 4\delta z$.

By analyzing the EMC and IMC schemes with these assumptions a linearized analysis can be done, and from such analysis it is possible to obtain some insights of the non-linear stability properties. The reasoning and manipulation can be seen in all extent in Appendix B. The stability condition for the EMC scheme is found to be

$$\Delta t \leq \frac{2\delta z}{\pi \nu_*} \left( \frac{1}{1 + \epsilon_*} \right) \tag{2.44}$$

where $\nu_* = \dfrac{K_o}{C}$ with viscosity units $[L^2/T]$ and the dimensionless number $\epsilon_* = \dfrac{3K_* b}{K_o}$. Equation (2.44) shows the charactestic form of a stability condition of a diffusion equation [1] with different coefficients, but nevertheless proportional to the square of mesh resolution, and inversely proportional to a viscosity coefficient. It is important to note that for $\epsilon_* \ll 1$ is less restrictive than equation (2.44). Note that the slope of the conductivity function, $K_*$,

participates only in $\epsilon_*$. Hence, it is only when $\epsilon_*$ is in the order of magnitude of 1 that the $K_*$ affects stability. High values of $K_*$ occur near saturation, specially if using the MG model. Hence, the MMG model is better suited to ensure stability than the MG model. This is consistent with the work by Schaap and van Genuchten [33] and Vogel et al. [39]. The viscocity coefficient $\nu_*$ is dependent on $C$. This allows a simple conclusion. Whenever very dry conditions, or saturation conditions arise, $C \to 0$, which implies $\nu_* \to \infty \Rightarrow \Delta t \to 0$. In order to further understand the stability properties of EMC to soil parameters, consider the Brooks-Corey model [9] and the Brutsaert equation for conductivity [11]. Then, stability can be written as

$$\Delta t = -\frac{2\delta z^2}{\pi} \frac{\omega(\theta_s - \theta_r)^2}{K_s h_b (\theta_o - \theta_r)} \left(\frac{\theta_s - \theta_r}{\theta_o - \theta_r}\right)^{\frac{3}{2\omega}} \left[\frac{1}{1 + 3b\left(2 + \frac{5}{2\omega}\right)\frac{(\theta_s - \theta_r)^2}{\theta_o - \theta_r}}\right] \tag{2.45}$$

where the $h_b$ is the *bubbling pressure* (similar to $h_s$ in MMG) and $\omega$ a fitting parameter that represents pore-size distribution (similar to $\hat{\eta}$ in MG and MMG).

From (2.45) it can be concluded that for a particular soil, $\Delta t$ is inversely proportional to saturation. The higher $\omega$ the more sensible $\Delta t$ is to saturation. Conversely, for a particular water content, there is a minimum value of $\Delta t$ for a particular $\omega$. The more saturated the soil is, the least sensitive $\Delta t$ is to $\omega$. Saturated conditions result in a $\Delta t$ which varies with $\omega$ very little around the minimum value of $\Delta t$. Hence, saturated, fine-textured soils are very restrictive on time step selection.

The analysis of the IMC leads to the conclusion that it is unconditionally stable. This must be considered in context, remembering the linearized analysis from where it derives. Additionally, because the stability analysis was approximated by using $K^n$, and not $K^{n+1}$, a significant part of the non-linearity of the problem might have been lost, specially since an arithmetic mean to obtain $K_{i\pm1}$ results in $K_{i\pm1} = K_o$.

### 2.4.5 Efficiency

A priori analysis of the schemes might lead to the conclusion that explicit schemes require less CPU-time since no iterations are necessary. However, because of stability constrains, if the admissible time step is small, CPU time can be much greater than the required CPU time for the implicit schemes. Since efficiency depends on stability constrains, then, the same factors which might lead to small time steps for the EMC scheme affect negatively on EMC efficiency.

IMC efficiency depends on the selected time step, but also on soil parameters and water content states. CPU-time requirements for the IMC are dependent on the number of iterations that need to be performed in each time step, times the number of time steps. Larger time steps require more iterations, the question is how many more iterations. An optimal time step might be sought to maximize efficiency. Because iterations intend to linearize $K(h)$, large gradients of $h$ and very non-linear conductivity functions will require more iterations, and hence efficiency is reduced. Nevertheless, it is only through simulation that it can be quantified. Furthermore, time step is likely to be selected according to the desired accuracy and compromising some diffusivity effects, hence, such optimal time step is not investigated in this work. Another factor in efficiency is the efficiency of the algebraic solution of the matrix equation. Because the schemes in this work are only 1D, high efficiency is achieved because of the Thomas Algorithm. However, in more dimensions, the matrix equation is not

tridiagonal and efficiency is likely to be severly affected. For such cases, appropriate selection of the alegebraic solver is essential.

## 2.5 Computation of intercell conductivity $K_{i\pm1/2}$

The issue of selecting an appropriate method to compute the intercell hydraulic conductivity (interblock conductivity, intergrid conductivity, or internode conductivity) has been extensively discussed in the literature and has been identified as a matter of great importance [8], as it can not only affect the quality of the results, but the stability of the numerical model [10]. Several schemes to compute the intercell conductivity have been proposed, analyzed and compared.

To illustrate the importance of the method for computing $K_{i\pm1/2}$, consider a discrete domain with constant and small $\delta z$. Whenever $\frac{\partial h}{\partial z}$ is small between two cells $i$ and $i+1$, the choice of an estimation method for $K_{i+1/2}$ should not be problematic, as the value will be tightly bounded by $K_i$ and $K_{i+1}$ which should be quite similar because of the small difference in pressure (this, however, has been shown not to be true in all cases [3]). Nevertheless, as the gradient of $h$ becomes larger between two cells, the estimation method becomes important. An inappropriate method, together with the non-linearity of $K(h)$ can lead to large misestimation of $K$ at the cell interface, thus errors in flow occur. This effect is magnified near saturation, as $\frac{\partial K}{\partial h} \to \infty$. This is especially important when solving a problem with boundary conditions that can generate large gradients near the boundaries, because of extreme fixed matric potentials. As fine grids become impractical for large scale or even catchment scale problems, the intercell conductivity estimation method needs to be robust enough to work with relatively coarse grids. The problem of computing intercell conductivity further extends to the estimation of interlayer conductivity in heterogeneous soils [10] [14] [31] and saturated-unsaturated interfaces [27].

Perhaps the most basic method, is the arithmetic mean,

$$K_{i\pm1/2} = \frac{K_i + K_{i\pm1}}{2} \tag{2.46}$$

The geometric mean has also been proposed,

$$K_{i\pm1/2} = \sqrt{K_i K_{i\pm1}} \tag{2.47}$$

The harmonic mean,

$$K_{i\pm1/2} = \frac{2}{\frac{1}{K_i} + \frac{1}{K_{i\pm1}}} \tag{2.48}$$

The upstream mean

$$K_{i\pm1/2} = \begin{cases} Max(K_i, K_{i\pm1}) & \text{if } \dfrac{\partial h}{\partial z} \geq 0 \\ \\ Min(K_i, K_{i\pm1}) & \text{if } \dfrac{\partial h}{\partial z} < 0 \end{cases} \tag{2.49}$$

Haverkamp and Vauclin [20] studied several methods and concluded that the geometric mean performs better than other methods. Hornung and Messing [21] showed that the geometric

mean performs better than the arithmetic mean. Zaidel and Russo [42] studied the Kirchhoff scheme as well as weighted methods relying on the asymptotic behavior of the conductivity function which for particular cases were reduced to a geometric mean. Van Dam and Feddes [36] used an arithmetic mean although it tends to overestimate infiltration rates (geometric means tend to underestimate it) but concluded that for fine grids the errors generated by the arithmetic mean are smaller than those produced by neglecting hysteresis and spatial soil variability. Gastó et al. [17] proposed a weighted averages method and found that the arithmetic mean overestimates and the geometric mean underestimates conductivity. Srivastava and Guzman [35] found that integrated conductivity (analitically or by Gaussian integration) provided good results, and also confirmed that the geometric mean outshines the arithmetic and harmonic means, but sometimes even the Gaussian integration method. Another interesting observation is that the upstream conductivity scheme and the harmonic mean scheme provide upper and lower boundaries of the exact solution. Belfort and Lehmann [8] performed simulations with several methods, validating the preference of the geometric mean over the arithmetic and harmonic mean (in particular for large $\delta z$), and also finding that for finite elements the geometric mean provides good results and efficieny, but for finite differences, weighted averages can prove better. They concluded that for large nodal spacing arithmetic and upstream means overestimate the wetting front and harmonic and downstream means underestimate it. Vanderborght et al. [38] conducted a set of benchmarking test cases among several codes, concluding that those which use the arithmetic mean predict more dispersed wetting fronts than those obtained with codes that use the geometric mean. Only one code obtained more dispersed fronts with the upstream mean.

Warrick [40] who also noted the arithmetic mean and even the geometric mean to be poor estimations, as the geometric mean in some cases greatly underestimated flow. Warrick proposed a weighting scheme which was found to be more accurate but also greatly increased computation time. More recently Baker [2] [3] further analyzed the validity of different means, evaluating if they satisfied mathematical principles (min-max conditions for elliptical value problems) and Darcian flow, and proposed a Darcian mean, i.e., a weighted mean obtained from the spatial distribution of $h$ which guarantees darcian flows, finding better accuracy than the geometric mean, but also noted the large computational overhead it requires. Baker showed that by comparing an analytical form of the Darcian mean using the Brooks-Corey model [9], this mean could be reduced to arithmetic, harmonic and geometric means depending on soil parameters, concluding that the arithmetic mean is representative of a nonphysical porous medium, the harmonic relates to an unlikely medium and the geometric mean to clays or rock matrices. Baker's results show that only the upstream mean did not violate mathematical principles, whilst the arithmetic, harmonic and geometric means showed a great number of violations. Furthermore, Baker showed that traditional means caused non-physical results, which did not occur with the upstream or Darcian mean, and that the main difference between the latter is that the Darcian mean produces sharper wetting fronts and higher peak flows than the upstream mean when space discretization errors occur. Baker's recommendation is that the upstream mean, because of computer efficiency, is in many cases preferable over the CPU-time-consuming Darcian mean.

Because of its simplicity and widespread use, despite the aforementioned studies, the arithmetic mean is included in the model. Nevertheless, following Bakers' recommendation [3], the upstream mean has also been included, as shown in equation (2.49). A simple way to understand the effects of choosing the arithmetic or upstream mean is to consider the soil shown in Figure 1.3. Consider a downward saturation process, for a cell interface between a saturated cell $i = 2$ ($K \rightarrow K_s$) and a dry boundary cell $i = 1$ ($K \rightarrow 0$) with a known and imposed pressure, estimation of the intercell conductivity with an arithmetic mean will result in $K_{1+1/2} \rightarrow K_s/2$. If the estimation of the intercell conductivity is done by using the

maximum value of conductivity of the adjacent cells (which corresponds in this case to an usptream mean), then $K_{1+1/2} \to K_s$. Because flow grows with $K$, the arithmetic mean will result in a lower downward flow (from cell 2 into cell 1) compared to using the maximum value of $K$, hence water content and pressure in cell 2 will increase faster than its surroundings, but water content in cell 1 remains unchanged, resulting in an unrealistic pressure gradient.

In a downward drying process the pressure gradient is negative, and thus all the above mentioned behaviors are inverse. In this case the use of the arithmetic mean wil result in $K_{1+1/2} \to K_s/2$ and the use of the minimum value will result in $K_{1+1/2} \to 0$. In terms of flow, this means that intercell flow is greater when using the arithmetic mean, thus, cell 2 will dry unrealistically faster while cell 1 remains with a constant water content.

## 2.6 Mass Balance Error Assesment

Mass balance in the domain, for any time $n\Delta t$ can be assessed by

$$\epsilon_{mb} = \frac{\overbrace{\sum_{i=1}^{N}\left(\theta_i^n - \theta_i^0\right)\delta z}^{\text{Change in mass}} - \overbrace{\sum_{j=1}^{n}\left(J_{1-1/2}^j - J_{N+1/2}^j\right)\Delta t}^{\text{Net flow}}}{\underbrace{\sum_{i=1}^{N}\theta_i^0\delta z}_{\text{Initial mass}}} \tag{2.50}$$

which represents the error in mass balance. Thus, when $\epsilon_{mb} = 0$ perfect mass balance is obtained. Note that net flow is equal to flow entering the domain from below ($J_{1-1/2} > 0$) minus flow exiting the domain in the upper boundary ($J_{N+1/2} > 0$). If flow is occuring downwards, the signs are inversed and the equation is still valid.

Note that equation (2.50) is relative to the initial mass in the domain which amplifies errors when the soil is initially in a very dry state, and reduces errors when the soil is initially saturated. Likewise, $\epsilon_{mb}$ can be expressed relative to the final mass in the domain ($\epsilon_{mbF}$), which would reduce the relative error in a saturation process, and amplify it in a desaturation process.

The computation of fluxes $J$ in (2.50) depends on the boundary conditions. For Dirichlet type (known matric potential) boundary conditions:

$$J_{N+1/2} = J_{N-1/2} = -K_{N-1/2}^j\left(\frac{h_N^j - h_{N-1}^j}{\delta z} + 1\right) \tag{2.51}$$

$$J_{1-1/2} = J_{1+1/2} = -K_{1+1/2}^j\left(\frac{h_2^j - h_1^j}{\delta z} + 1\right) \tag{2.52}$$

Note that when the lower boundary condition is set to gravity flow, equation (2.52) is also valid, with the particularity that $(h_2 - h_1)/\delta z = 0$. For Neumann type (known flow and impervious stratum) boundary conditions:

$$J_{N+1/2} = J_* \tag{2.53}$$

$$J_{1-1/2} = J_o \tag{2.54}$$

# Chapter 3

# Validation and test cases

## 3.1 Warrick's Analytical Solution

In order to test the validity of the model, comparison against an analytical solution is first performed. Because the MG model has been selected for consitutive relations of the soil, an analytical solution which makes use of such model is preferred. Warrick et. al. [41] proposed a generalized solution for an infiltration problem which has been used as analytical benchmarking in [28].

Warrick's solution considers a dimensionless form of the water content Richards' equation.

$$\frac{\partial W}{\partial T} = \frac{\partial}{\partial X}\left(D^* \frac{\partial W}{\partial X} + K^*\right) \tag{3.1}$$

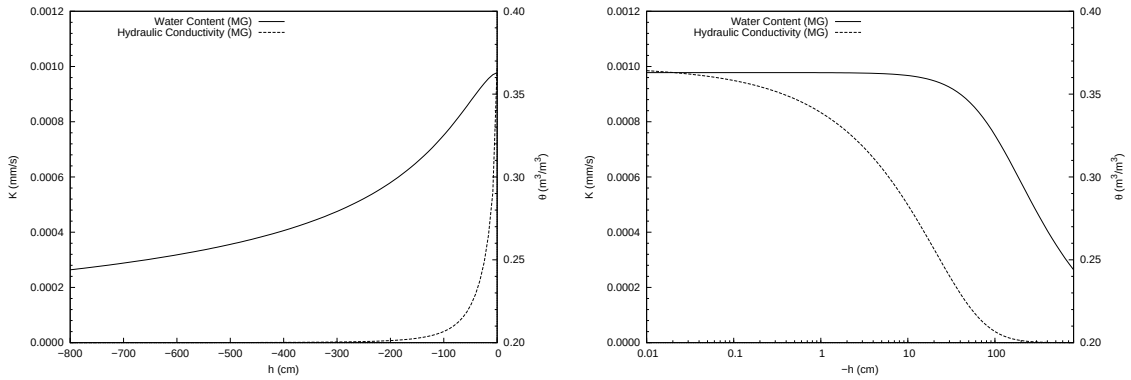
where the dimensionless variables are $W = \left(\frac{\theta - \theta_r}{\theta_s - \theta_r}\right)$, $T = \frac{\alpha K_s t}{\theta_s \theta_r}$, $X = \alpha z$, $K^* = \frac{K}{K_s}$ and $D^* = K^* \frac{\partial(\alpha h)}{\partial W}$.

Considering a semi-infinite soil column (lower boundary condition) with an initial condition of constant (unsaturated) water content along the column, this is $h(z, t = 0) = h_0 < 0$ which implies $\theta(z, t = 0) = \theta_0$ and finally $W(X, T = 0) = W_0$. The upper boundary intends to simulate infiltration by imposing $h(x = L, t) = 0$. For such conditions, the solution is given by

$$X = \lambda(W) T^{\frac{1}{2}} + \chi(W) T + \psi(W) T^{\frac{3}{2}} \tag{3.2}$$

Coefficients $\lambda$, $\chi$ and $\psi$ are functions of $W$, $W_i$ and $\hat{\eta}$, and can be found in [28]. The solution algorithm is as follows. Values for the coefficients are known for certain values of $W$ which correspond to values of $\theta$. By using equation (3.2) and the dimensionless variables definitions, values of $\theta(z)$ can be obtained. By applying the MG model, $h(z)$ values are found.

To compare model results to the analytical solution, the following soil parameters for the MMG model were used: $\theta_s = 0.363$, $\theta_r = 0.186$, $\alpha = 0.01\ cm^{-1}$, $\hat{\eta} = 1.53$, $K_s = 0.0001\ cm/s$, $h_s = 0\ cm$. Soil curves are shown in figure 3.1(a) and in 3.1(b) in logarithmic scale for pressure for clarity. The initial condition was set to $h_0 = -800\ cm$ and the soil column depth was 100 $cm$.




(a) (b)

Figure 3.1: Soil properties for validation tests

Table 3.1 summarizes the validation tests. The effects of mesh size, time step and conductivity averaging technique are observed for the schemes. All schemes are tested against the analytical solution for two fine meshes. The effects of conductivity averaging are explored only for the IMC scheme since it is the most sensitive scheme to this parameter. Time step effects are also explored for the IMC scheme only, since the EP and IP scheme are shown to be inaccurate, and the EMC scheme is unstable for larger time steps than the one used for validation.

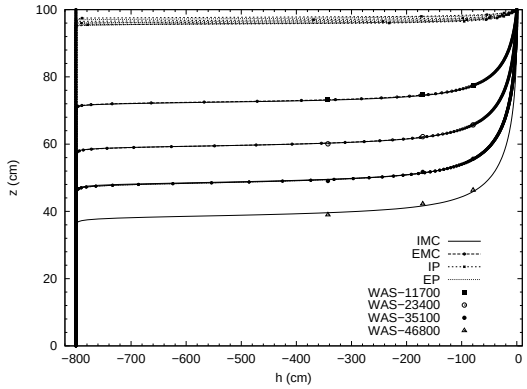
Table 3.1: Validation Tests

| | | | | $\Delta t\,[s]$ | | | |
|---|---|---|---|---|---|---|---|
| Test | Figure | Averaging | $\delta z\,[cm]$ | EP | EMC | IP | IMC |
| 1 | 3.2(a) | A | 0.25 | 0.001 | 0.0001 | 1 | 1 |
| 2 | 3.2(b) | A | 1 | 0.001 | 0.0001 | 1 | 1 |
| 3 | 3.2(c) | A | 1 | - | - | - | 1-11700 |
| 4 | 3.2(d) | A | 1, 5, 20 | - | - | - | 1 |
| 5 | 3.4(a) | A, US | 0.25 | - | 0.001 | - | 1 |
| 6 | 3.4(b) | A, US | 1 | - | 0.001 | - | 1 |
| 7 | 3.4(c) | A, US | 0.25 | - | - | - | 780 |
| 8 | 3.4(d) | A, US | 1 | - | - | - | 780 |

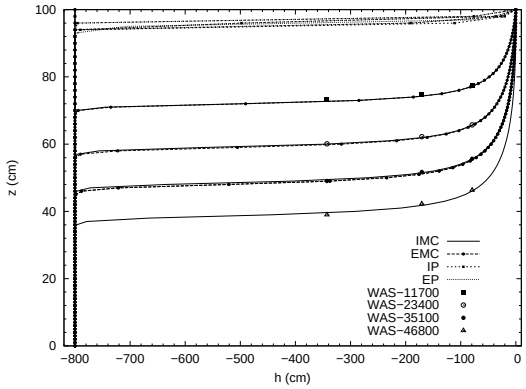
A comparison between the solution obtained with each scheme against the analytical solution for a mesh of $\delta z = 0.25\,cm$ is shown in figure 3.2(b). For IMC $\Delta t = 1\,s$. For EMC, $\Delta t = 0.0001\,s$. However instabilities appeared by the end of the simulation and $t = 46800\,s$ could not be computed. Using the stability criterion, equation (2.44), for the wetting "perturbation" results in $\Delta t_{max} \approx 8\times10^{-6}$. Tests with $\Delta t = 1\times10^{-5}$ were performed which became unstable far into the simulation. Because of the enormous simulation time required (a week by the time the instability appeared), no further tests were performed. CPU time for IMC was approximately 26 hours long. For EMC, CPU time was almost 41 hours up to the moment of instability (around $t = 42000\,s$). Note that for EMC no results are presented for $t = 46800\,s$. The results show that both IMC and EMC schemes are capable of accurately approximating the solution, considering the range in which EMC was stable. The differences between EMC and IMC are negligible in both meshes. EP and IP schemes are severely inaccurate, although EP remained stable. Note the similarity in the erroneous results of both IP and EP.

A comparison between the solution obtained with each scheme against the analytical solution for a mesh of $\delta z = 1\,cm$ is shown in figure 3.2(b). For IMC $\Delta t = 1\,s$. For EMC, from

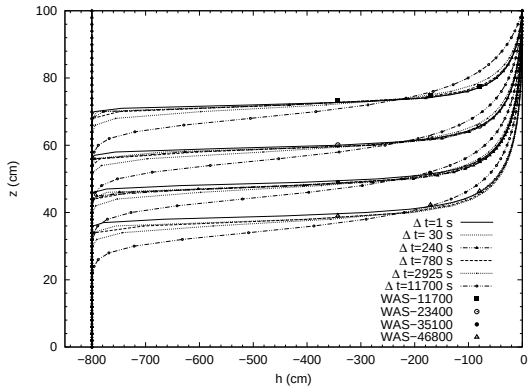
equation (2.44) for the wetting "perturbation", $\Delta t_{max} \approx 3 \times 10^{-5}$, although with $\Delta t = 0.0001\ s$ was sufficient to obtain results for the entire simulation time. CPU time for IMC was approximately 5.7 hours. For EMC, CPU time was around 18 hours. The results show that both schemes are capable of accurately approximating the solution and produce practically the same solution. EP and IP schemes are severely inaccurate, although EP remained stable. Note the similarity in the erroneous results of both IP and EP.



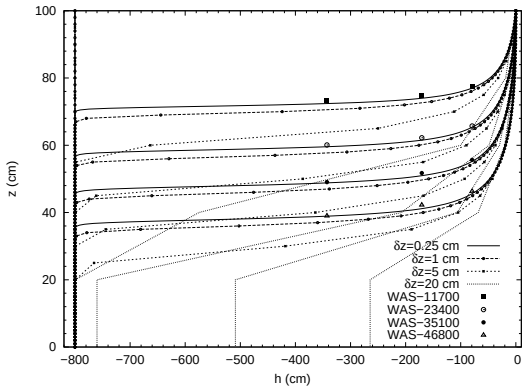

(a) Scheme comparison, $\delta z = 0.25$, arithmetic mean




(b) Scheme comparison, $\delta z = 1$, arithmetic mean




(c) Effect of $\Delta t$ in IMC, arithmetic mean, $\delta z = 1$



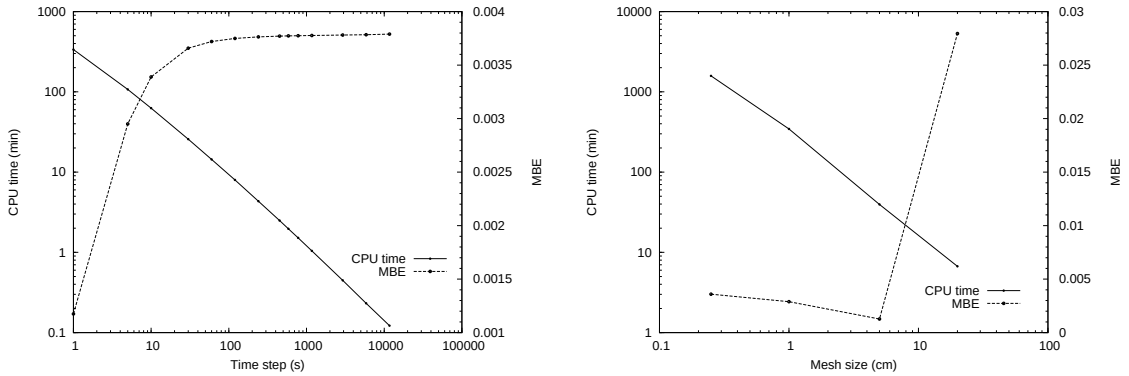

(d) Effect of mesh size. IMC, $\Delta t = 1$.

Figure 3.2: Simulation results compared with Warrick's analytical solution

Because the two basic tests show that EP and IP are inaccurate, they were no longer used in further tests.

The same setup was used to test the effects of $\Delta t$ for the IMC scheme. The EMC scheme was not tested since the effect of $\Delta t > 0.0001$ resulted in unstable behavior. The effects on the solution are shown in 3.2(c). The position of the wetting front is in general well described even with large time steps. The effect of larger $\Delta t$ is that the wetting front rotates slightly and is smoother, more diffused. The effects over CPU time and mass balance error are shown in figure 3.3(a) (note that this curves were constructed with more time steps than those shown in figure 3.2(c). Note that CPU time decreases rapidly for larger time steps. MBE grows for large values of $\Delta t$ but stabilizes asymptotically to a maximum value. Under relaxation techinques were also tested for several $\Delta t$ but no significant differences in CPU time were found.

The effect of mesh size can be seen by comparing figures 3.2(a) and 3.2(b). The finest mesh of $\delta z = 0.25$ results in better accuracy. The front advances slightly faster with a coarser mesh for both the EMC and IMC schemes. The difference between EMC and IMC, for each mesh are neglegible. In the coarser mesh differences between both schemes are slightly larger, but




(a) Effect of $\Delta T$

(b) Effect of $\delta z$

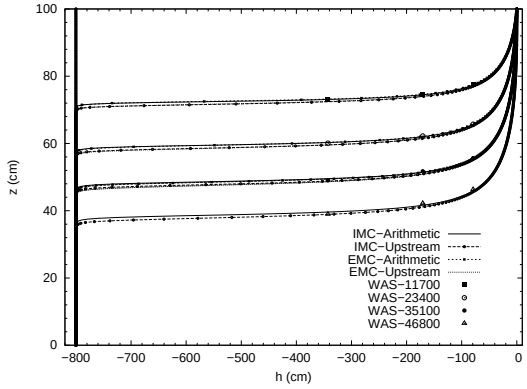
Figure 3.3: CPU time for IMC

nevertheless small.

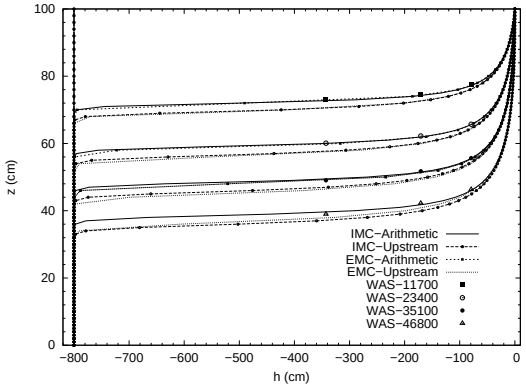
Furthermore, figure 3.2(d) shows results of the IMC scheme for the same time step for four different meshes. It is clear that the finest mesh yields the most accurate result. A slightly coarser mesh still yields accurate results for the shape and position of the wetting front. For $\delta z = 5$ the front is much faster, and for $\delta z = 20$ the shape and position of the front is almost lost. Figure 3.3(b) shows the decrease of CPU time as a function of mesh size. Although a reduction in CPU time exists, it is not as efficient or accurate to use coarser grids as it is to use larger time steps with the IMC scheme.

Comparison of figures 3.4(a) and 3.4(b) shows that by selecting different conductivity averaging techniques slight differences occur in both the IMC and EMC schemes. The upstream mean generates a slightly faster wetting front for both schemes. The effect of the averaging technique is smaller for finer grids, as was expected (see section 2.5), since averaging is in fact interpolating values. Hence a smaller grid is less senstive to interpolation errors. These figures also show that the EMC solution and the IMC solution are practically identical with the same averaging technique for each mesh. In the coarser mesh there are slightly more differences, with a tendency for EMC with upstream mean to generate the fastest front, and IMC with arithmetic mean to generate the slowest front. Nevertheless, there are more differences in accuracy by selecting the averaging technique than by using EMC or IMC.

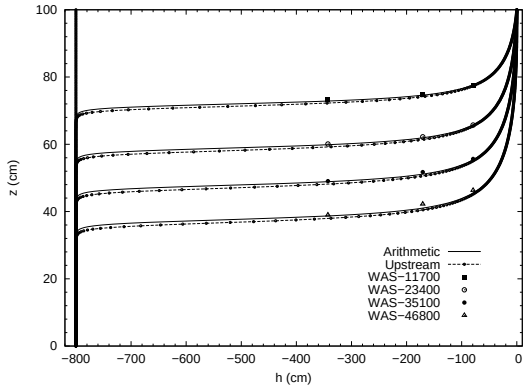
Figures 3.4(c) and 3.4(d) show the same comparison with a dramatically larger time step which was proven to be accurate in figure 3.2(c). Only IMC scheme solutions were computed, since EMC allows for little time step manipulation. Note that $\Delta t$ does not affect the way the averaging technique. Although the solution is not the same as for $\Delta t = 1$, there is no apreciable amplification of the variations introduced by different averaging techniques. Only mesh selection affects the behavior of the averaging technique.



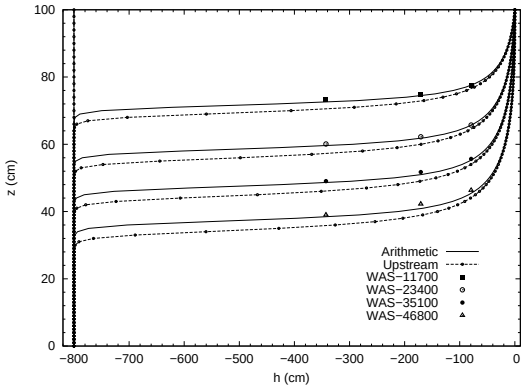

(a) EMC, IMC $\Delta t = 1$, $\delta z = 0.25$




(b) EMC, IMC $\Delta t = 1$, $\delta z = 1$




(c) IMC, $\Delta t = 780$, $\delta z = 0.25$




(d) IMC, $\Delta t = 780$, $\delta z = 1$

Figure 3.4: Effects of conductivity averaging compared with Warrick's analytical solution

## 3.2 Test cases

A series of test cases were simulated, considering a soil stratum of 100 $cm$ in depth with the following parameters for the MG and MMG models: $K_s = 0.00922\ cm/s$, $\theta_s = 0.368\ m^3/m^3$, $\theta_r = 0.102\ m^3/m^3$, $\alpha = 0.0335\ cm^{-1}$, $\hat{\eta} = 2$ (these parameters generate $K$ and $\theta$ curves as shown in figure 1.3). For MMG, $h_s = -4\,cm$. For all cases a uniform fine grid of $\Delta t = 1\ s$, $\delta z = 1\ cm$ was kept as a standard for comparison, and convergence criteria $\epsilon = 10^{-7}$, unless otherwise noted. This is considered to be very strict, in comparison to [22]. For every test case, four simulations were performed to observe sensitivity to particular methods. All test cases are downward processes. A summary of the test cases is presented in table 3.2, and a summary of the setup of the different simulations for each case is presented in table 3.3. Note that only the IMC and EMC schemes were used for these test cases, since the IP and EP schemes were proven inaccurate in validation tests in section 3.1. The EMC scheme was used in those cases in which saturation conditions need not be computed, since the scheme is ineffective when saturation conditions arise. Although cases 6 and 7 appear to be well suited for EMC, when tested, the EMC scheme either was not capable of advancing the drying front (because of initial saturation conditions) or became unstable.

Table 3.2: Test Cases

| Test Case | Description | Scheme | UBC | LBC | IC |
|---|---|---|---|---|---|
| 1 | Impervious boundaries | IMC | $J_* = 0\,cm/s$ | $J_o = 0\,cm/s$ | $h = -20\,cm$ |
| 2 | Saturation in semi-infinite soil | IMC | $h = 0\,cm$ | $\frac{\partial h}{\partial z} = 0$ | $h = -100\,cm$ |
| 3 | Saturation with water table | IMC | $h = 0\,cm\,cm$ | $h = 0\,cm$ | $h = -100\,cm$ |
| 4 | Partial saturation | IMC-EMC | $h = -20\,cm$ | $h = -50\,cm$ | $h = -50\,cm$ |
| 5 | Full saturation | IMC | $h = 0\,cm$ | $h = -50\,cm$ | $h = -50\,cm$ |
| 6 | Drying with water table | IMC | $h = -100\,cm$ | $h = 0\,cm$ | $h = 0\,cm$ |
| 7 | Drying with semi-infinite soil | IMC | $J_* = 0\,cm/s$ | $\frac{\partial h}{\partial z} = 0$ | $h = 0\,cm$ |

UBC: Upper boundary condition; LBC: Lower boundary condition; IC: Initial condition

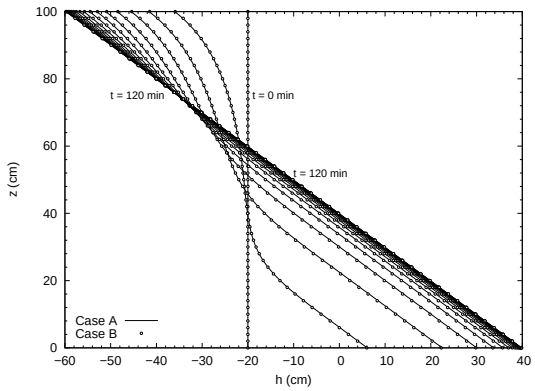
Table 3.3: Simulation setup

| Simulation case | Mean | Soil Model |
|---|---|---|
| A | Arithmetic | Mualem-van Genuchten |
| B | Upstream | Mualem-van Genuchten |
| C | Arithmetic | Modified Mualem-van Genuchten |
| D | Upstream | Modified Mualem-van Genuchten |

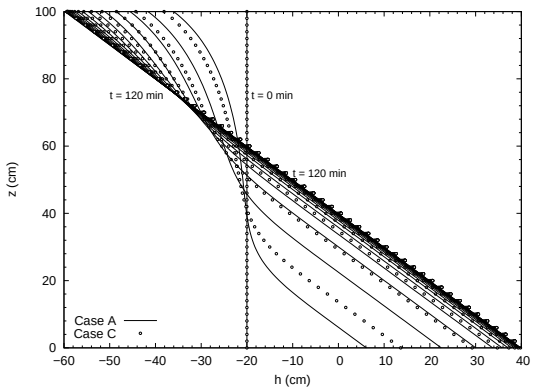
### 3.2.1 Test Case 1: Impervious boundaries

Consider a soil column overlying an impervious stratum ($J_o = 0\,cm/s$) and no infiltration from the surface ($J_* = 0\,cm/s$). Hence, the domain has zero net flow and there is no change in total mass within the soil column. With an initial state of a partially saturated column $h(z, t = 0) = -20\ cm$, the only changes should be the redistribution of water content because of gravity. The simulation was 2 hours long, with results shown every 5 minutes.

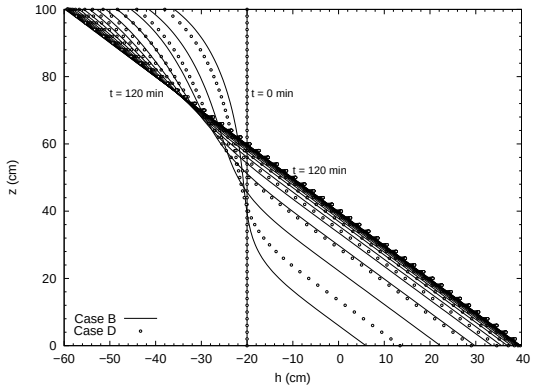
Simulation results are as expected. No flow enters or exits the domain, and perfect mass balance was obtained in all four simulations. The matric potential profile changes in time, as gravity forces water down, saturating the lower parts of the column, and drying the upper parts until equilibrium (zero flow) matric potential profile is obtained. Theoretically, from



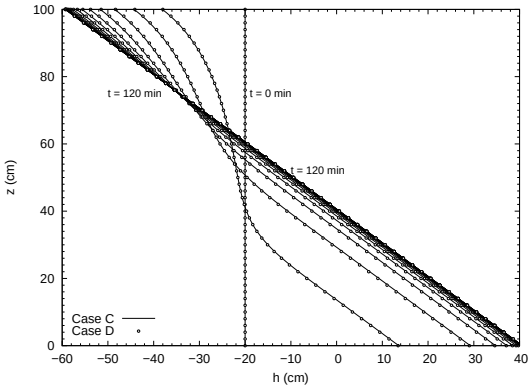

(a) Simulations A and B




(b) Simulations A and C




(c) Simulations B and D




(d) Simulations C and D

Figure 3.5: Results for Test Case 1

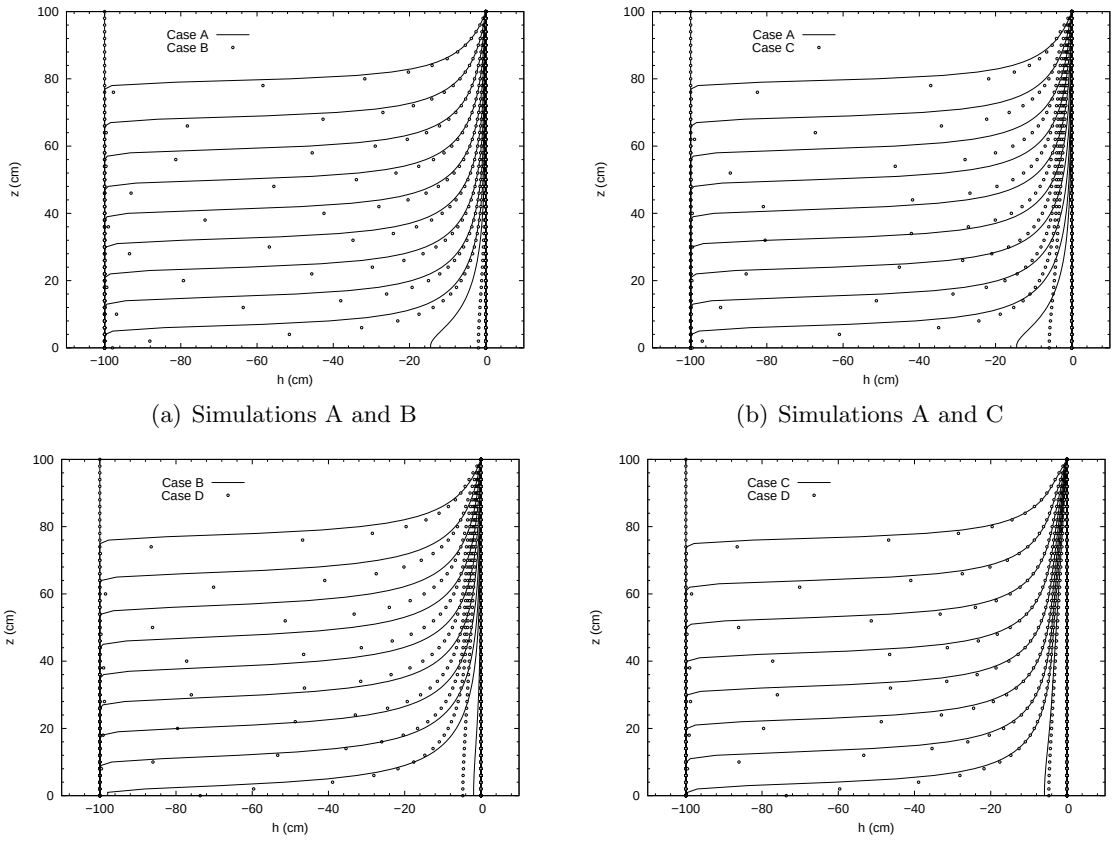
equation (1.40), it is clear that this equilibrium requires $\frac{\partial h}{\partial z} = -1$ so that gravitational potential is counteracted. This is verified in the simulation and is easily observed as the uniform slope in the final matric potential profile.

Although it is difficult to observe in figure 3.5(a) because of the scale, closer examination shows that in the lower parts of the column (wetting front) in case B, for the same time and depth, shows less hydrostatic pressure than case A. In other words, case B produces a slower wetting front than case A. Conversely, for the drying front in the upper part of the soil column, case B produces slower drying than case A. Nevertheless, as figure 3.5(a) shows, the difference is minimal. Similar behavior is found in 3.5(d). This shows that the upstream mean produces faster wetting fronts compared to the arithmetic mean.

Comparison between MG and MMG models (figures 3.5(b) and 3.5(c)) shows clearly that the MMG model generates a faster wetting front, which is due to the fact that maximum conductivity is achieved at lower water contents.

### 3.2.2 Test Case 2: Downward saturation in semi-infinite soil

Initial and boundary conditions were imagined so that the complete saturation process could be observed, allowing gravitational flow in the lower boundary which can be interpreted as having a semi-infinite stratum of the same soil. Initial conditions were $h(x, t = 0) = -100\,cm$. Boundary conditions were $\frac{\partial h}{\partial z}|_{x=0,t} = 0$ and $h(x = 100, t) = 0\,cm$. The simulation was 1 hour long, with results shown every 3 minutes.




(a) Simulations A and B

(b) Simulations A and C

(c) Simulations B and D

(d) Simulations C and D

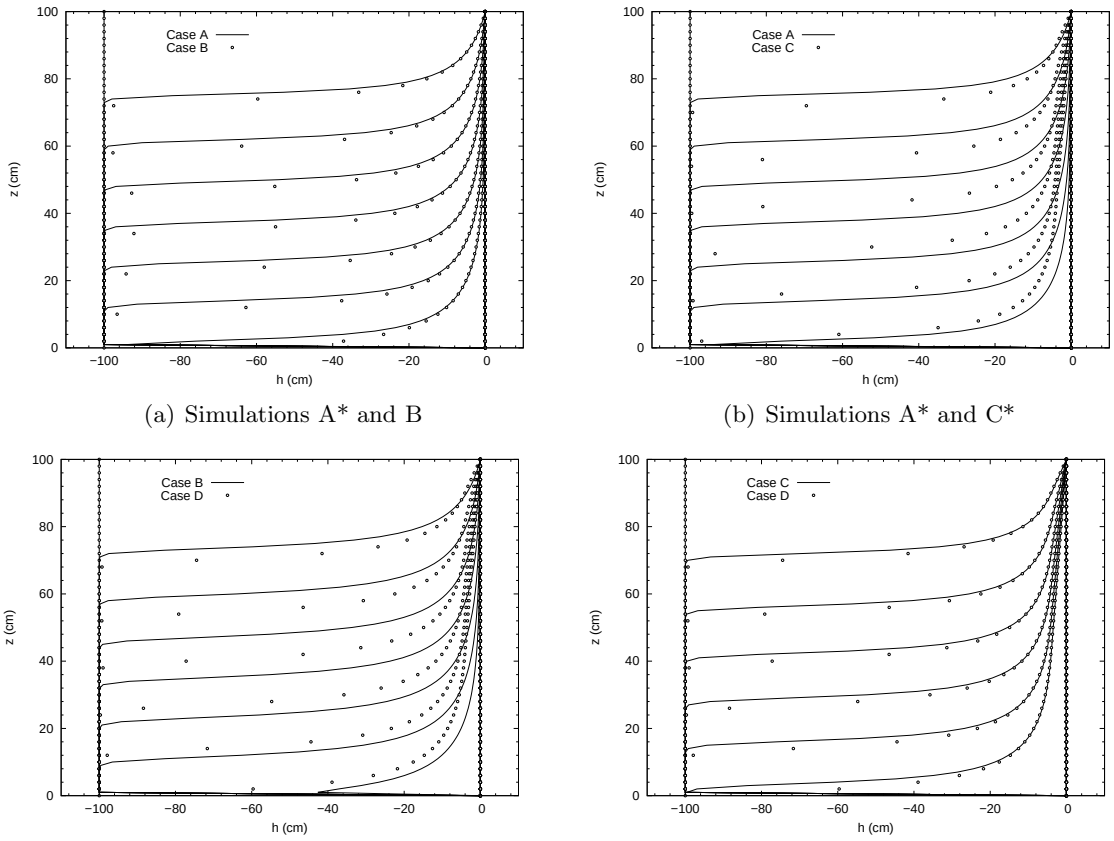
Figure 3.6: Results for Test Case 2

Note that although the entire column becomes saturated, pressure head never becomes positive. This is due to the free outflow at the lower boundary by setting the pressure gradient as zero. The effects of the conductivity mean are consistent with those seen in the validation test, showing faster fronts with upstream means. The soil constitutive model generates larger differences in the advancement of the front than mean selection, MMG generates faster fronts than MG.

### 3.2.3 Test Case 3: Downward saturation with water table

This case models a complete saturation process with presence of a fixed water table at the lower boundary. Initial conditions were $h(z, t = 0) = -100\,cm$. Boundary conditions were $h(z = 0, t) = 0\,cm$ and $h(z = 100, t) = 0\,cm$. The simulation was 1 hour long, with results shown every 4 minutes. Simulation for Case A became unstable after 20 seconds, and thus no results are shown for clarity, and because of the same reason no results are shown for case C.

This case has positive and negative pressure head gradients, which generate a particular scenario for the estimation method for $K_{i\pm1/2}$. The complications are evident in the overshooting effects near the water table boundary, which explain why Simulation A and C failed. Cases B and D do not show overshooting effects, although the negative gradient near the bottom still exists, thus, the arithmetic mean is responsible for the overshooting effects. An additional simulation was performed considering an automatic selection of the mean, in such a way that computation is performed with the arithmetic mean if there is no change in the sign of the




(a) Simulations A* and B

(b) Simulations A* and C*

(c) Simulations B and D

(d) Simulations C* and D

Figure 3.7: Results for Test Case 3

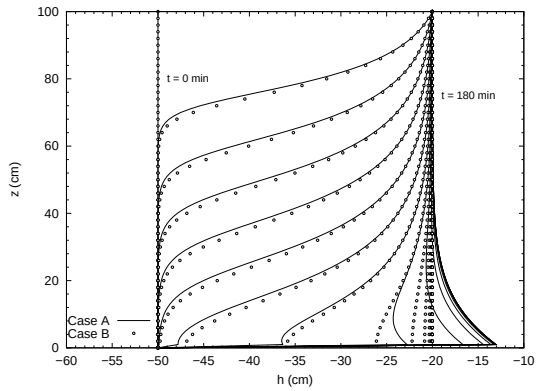
pressure gradient in the three-point finite difference stencil, and when there is a change in the sign, computation is performed with the upstream mean. Results are reported as case A* and C*.

Because cases A and C cannot be properly computed, it is not possible to compare the advance of the wetting front with cases A* and C*. However, it is interesting to note that case A* exhibits a slower wetting front than C* as it is expected. The use of the arithmetic mean as a primary method in cases A* and C* (and the upstream as a correction against overshooting) produces a slightly slower wetting front than cases B and D which are "fully" upstream.

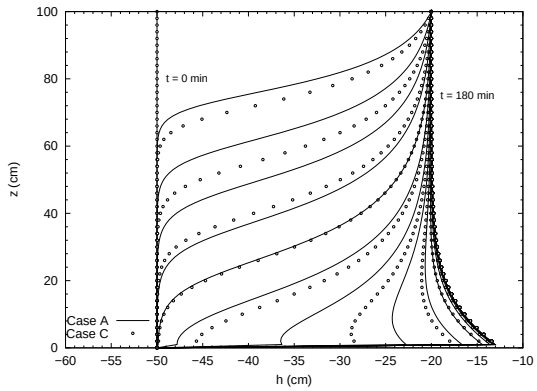
### 3.2.4 Test Case 4: Downward partial saturation process

This case models a saturation process which leads to a partially saturated stationary flow. Initial conditions were $h(x, t = 0) = -50\,cm$. Boundary conditions were $h(x = 0, t) = -50\,cm$ and $h(x = 100, t) = -20\,cm$. The simulation time was 3 hours long. Results are shown every 10 minutes. Note that stationary flow occurs around 110 minutes. For this case tests were performes also with the EMC scheme. Results shown are those of simulations with $\Delta t = 0.7\,s$.

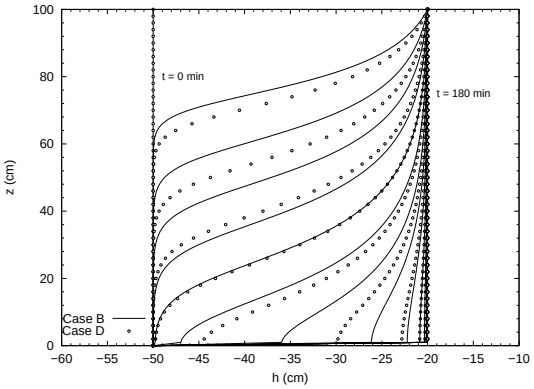
The overshooting effect generated by different means can be seen quite clearly. When the arithmetic mean is used, the intercell conductivity in the lower boundary is greatly affected by the unvarying and very low conductivity of the boundary, which in turn is overcompensated by accumulating mass (and hence pressure) in the column. When the upstream mean is



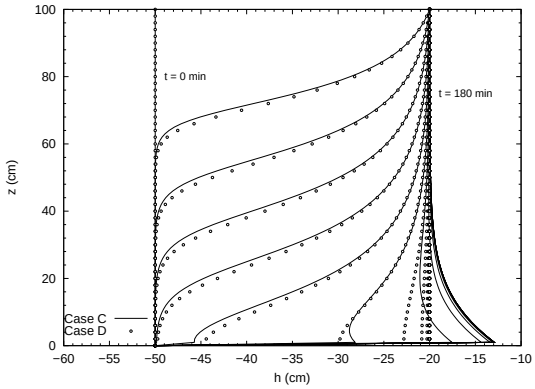

(a) Simulations A and B




(b) Simulations A and C




(c) Simulations B and D



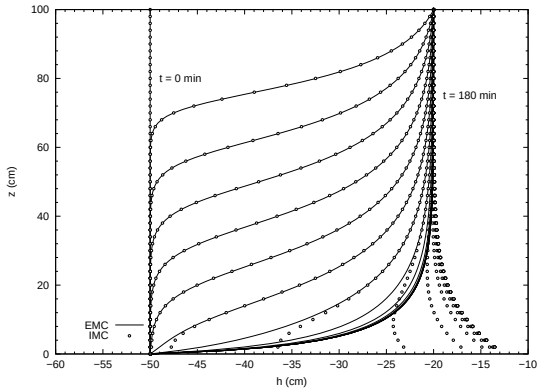

(d) Simulations C and D

Figure 3.8: Results for Test Case 4 with IMC

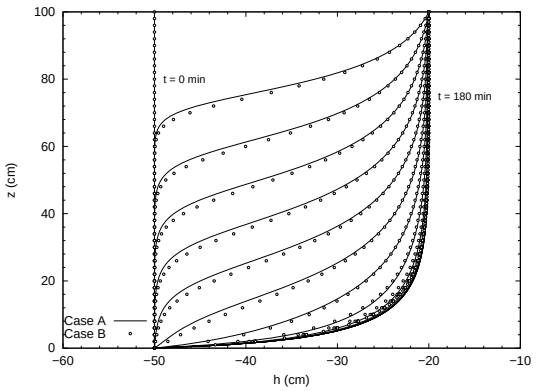
used, the lower boundary does not interact with the soil, hence, no overshooting occurs. Nevertheless, in spite of the overshooting effects, the speed of the wetting front is much more sensitive to changes in the consitutive model: MMG produces faster fronts than MG. The small effect produced by selecting upstream or arithmetic means in the speed of the wetting front can be seen in figures 3.8(a) and 3.8(d), which show that upstream produces faster fronts.

Figure (3.2.4) shows results for this case using the EMC scheme. The comparisons are the same as those with the IMC scheme, except for figure 3.9(a) which shows results of case A with EMC and IMC. Note that the solution is the same except in the boundary. EMC does not experience overshooting issues. The response of EMC to the use of arithmetic or upstream means is the same as for IMC: upstream means generate faster fronts. The use of MMG generates faster fronts than MG also. In terms of stability, $\Delta t > 0.7$ generated instabilities with the MMG model. With $\Delta t \approx 1$ instabilities appeared with the MG model. Evaluating equation (2.44) with $K_o(h = -20) = 0.0022067$, $C(h = -20) = 0.103$, $K_* = 0.06028$, with perturbation $b = 0.08538$ results in $\Delta t_{max} = 3.7$. Hence, consider the stability number

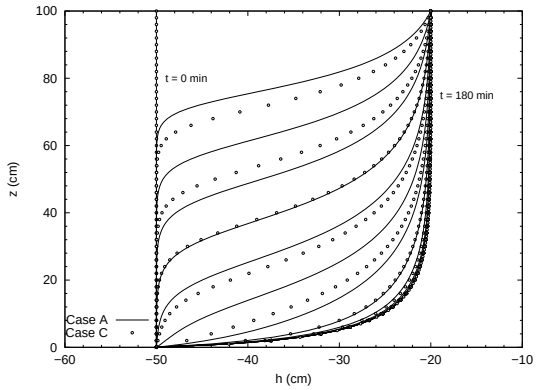
$$SN = \frac{\Delta t_{real}}{\Delta t_{max}} = \frac{0.7}{3.7} = 0.189$$



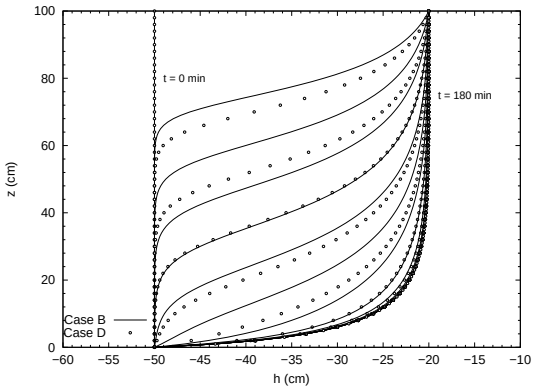

(a) Simulation A vs IMC



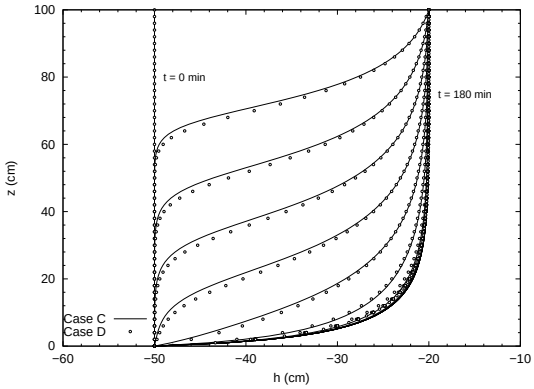

(b) Simulations A and B




(c) Simulations A and C




(d) Simulations B and D

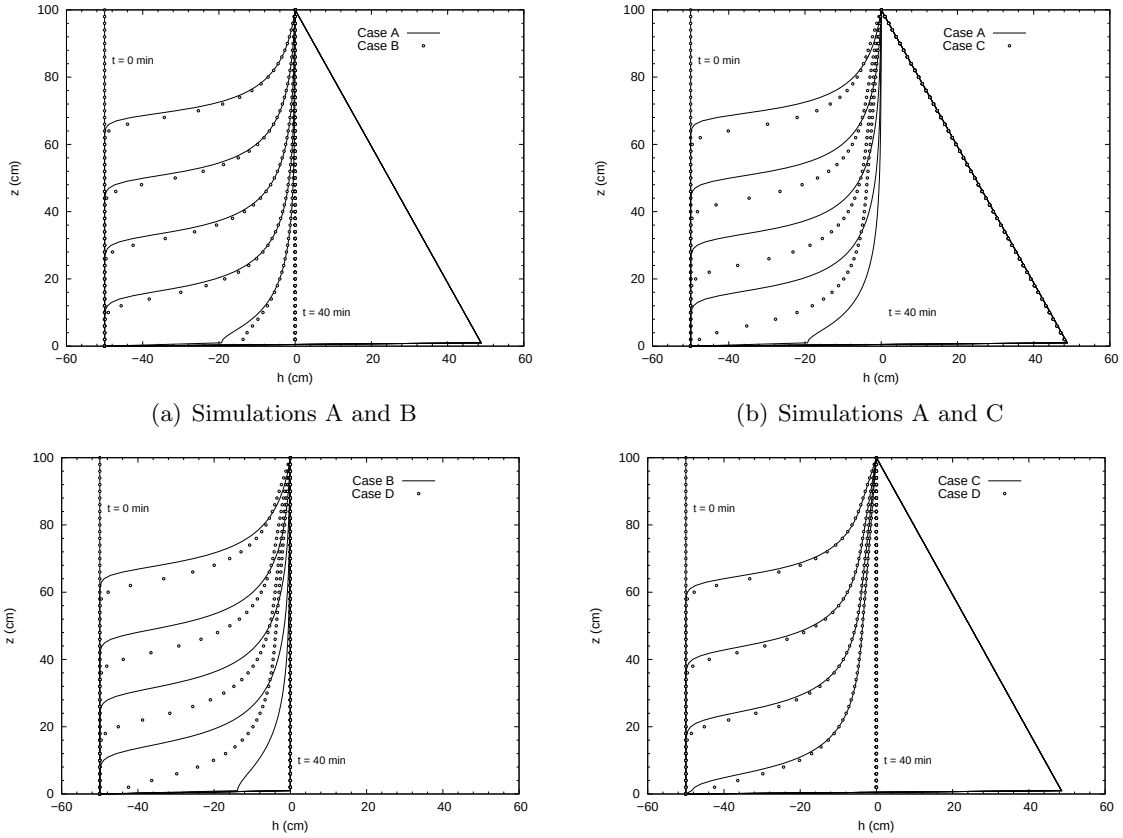



(e) Simulations C and D

Figure 3.9: Results for Test Case 4 with EMC

### 3.2.5 Test Case 5: Downward full saturation process

This case models the complete saturation process with an imposed pressure head in the lower boundary. Intial conditions were $h(x, t = 0) = -50\,cm$. Boundary conditions were $h(x = 0, t) = 0\,cm$ and $h(x = 100, t) = -50\,cm$. Simulation time was 2400 seconds (40 minutes) long. Results are shown every 240 seconds (4 minutes). Simulation B showed convergence issues near saturation, which required to set $\epsilon = 10^{-6}$. Note that stationary flow occured around 24 minutes into the simulation.




(a) Simulations A and B

(b) Simulations A and C

(c) Simulations B and D

(d) Simulations C and D

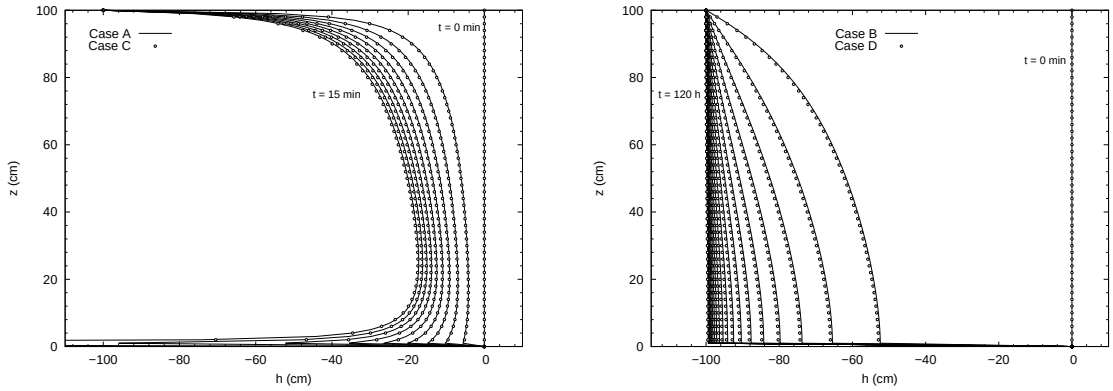
Figure 3.10: Results for Test Case 5

Results for this case are an extension of those of Case 4 into a saturated regime, because the upper boundary condition is set to induce saturation in the entire profile. Conclusions are quite similar to those of Case 4, but in this case it should be noted that the overshooting effect generates a blockage response of the lower boundary, resulting in a fictitious impervious boundary. This shows very clearly that, setting a Dirichlet condition in the lower boundary, and by doing so imposing a conductivity, can dramatically change the pressure profile. It should be noted that it was possible to fully simulate cases A and C, contrary to Case 3 and there was no need to use the automatic selection for averaging $K$. The effects of the MMG and MG model are as in previous cases, and have no incidence on overshooting which continues to be a conductivity averaging issue.

### 3.2.6 Test Case 6: Downward drying process with water table

This case models a drying process while maintaining a fixed water table in the lower boundary. Initial conditions were set as $h(x, t = 0) = 0\,cm$. Boundary conditions were $h(x = 0, t) = 0\,cm$

and $h(x = 100, t) = -100\,cm$. This simulation used $\Delta t = 10s$ and $\delta z = 2cm$. Because simulations A and C became unstable around $t = 15$ minutes, results are presented every 100 seconds. Simulations B and D were 120 hours long with results shown every 6 hours. Note that stationary flow is reached by the end of the simulation. Comparisons AB and CD provide no interesting information because of the difference in time scales, hence, they are omitted.




(a) Simulations A and C

(b) Simulations B and D

Figure 3.11: Results for Test Case 6

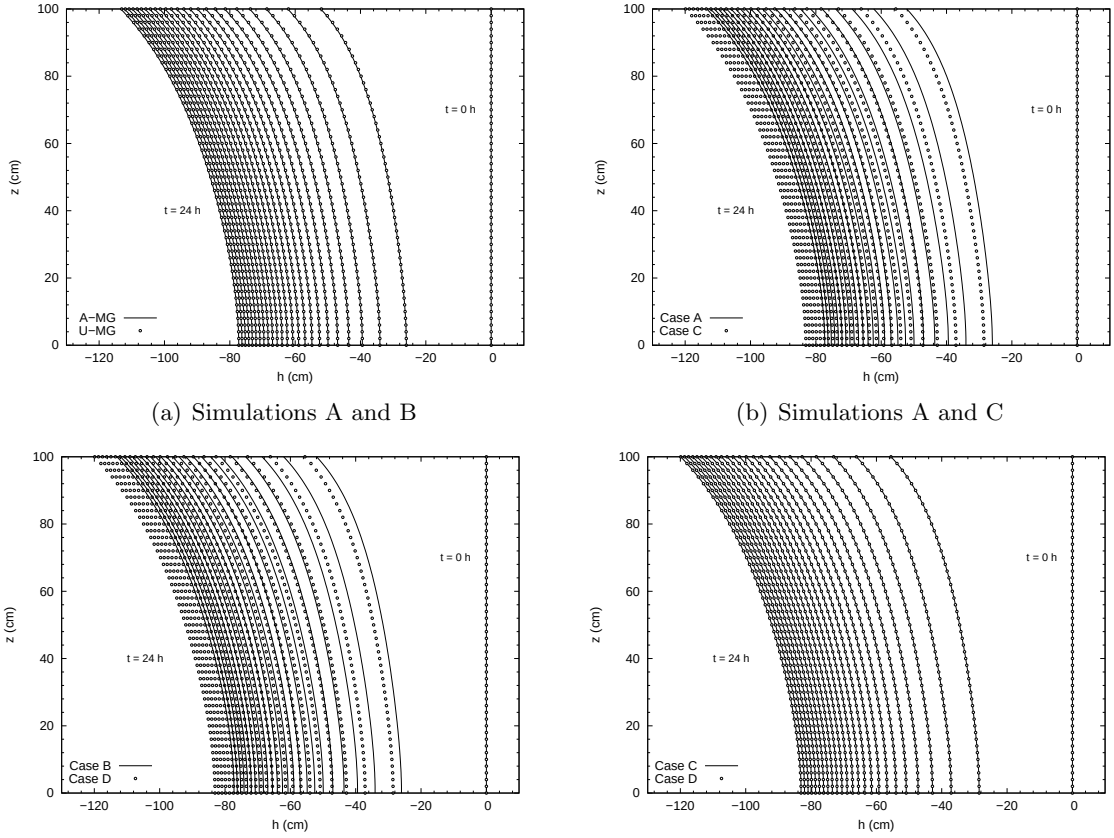
Figure 3.11(a) clearly shows the effect of the Dirichlet boundary condition together with the arithmetic mean. It seems clear that the "overdryness" of cell 2 in simulations A and C is the cause of the instability. Because simulations B and D do not show this behavior, it is an indication that the arithmetic mean does not handle well the Dirichlet-type boundary, resulting in overshooting, and eventually, instability.

In can be seen from the results that in this drying process, although it is the same soil as in the wetting cases, the use of MMG instead of MG generates faster fronts, but with very little difference in contrast to those generated by MG.

### 3.2.7 Test Case 7: Downward drying process in semi-infinite soil

This test case consists of an initially saturated soil column $h(z, t = 0)$=0 $cm$ of a semi-infinite soil. Boundary conditions were set as Neumann conditions: no flow from the surface $J_* = 0\,cm/s$ and a semi-infinite free draining stratum, which is represented by $\frac{\partial h}{\partial z}|_{x=0,t} = 0$. The simulation was 1 day long, with results shown every hour.

Because only Neumann boundary conditions are used, no overshooting issues arise when using the arithmetic mean. Furthermore, the differences generated by selecting the arithmetic mean or the upstream mean are negligible as shown in figures 3.12(a) and 3.12(d). There are far more significant differences when using the MMG or MG constitutive models as can be seen in figures 3.12(b) and 3.12(c), which show that the MMG model generates faster drying fronts. This is curious, since test case 6 resulted in very little differences with respect to both constitutive models.




(a) Simulations A and B

(b) Simulations A and C

(c) Simulations B and D

(d) Simulations C and D

Figure 3.12: Results for Test Case 7

# Chapter 4

# Conclusions and further research

## 4.1 Conclusions

1. The EP (Explicit, Pressure form of Richards' equation) and IP (Implicit, Pressure form of Richards equation) schemes do not converge to correct solutions. They both show very poor conservation properties.

2. The EMC (Explicit, Mixed form of Richards' equation) is mass conserative and accurate, although incapable of solving saturated conditions, and conditionally stable.

3. The IMC (Implicit, Mixed form of Richards' equation) scheme, formulated in a similar manner to a pressure form, in such a way as to solve pressure and not water content is appropriate to solve variably saturated flow, with adequate mass conservation, and reasonable efficiency.

4. EMC and IMC scheme approximate correctly the solution to Richards' equation. Differences in the solutions were only observed near Dirichlet boundaries, where de EMC scheme generates smooth transitions, while de IMC scheme results in discontinuities or even overshooting.

5. The EMC scheme can be computationally more efficient than IMC in certain cases (particularly for the same $\Delta t$). However, in other cases EMC can be severely less efficient because of stability constrains.

6. The stability analysis of the IMC scheme and the test cases show that the scheme is inconditially stable.

7. The stability analysis of the EMC scheme shows that it is conditionally stable, in a way similar to that of a diffusion equation. Valditation and tests cases confirm such dependence, although, because of the non-linearities of the equation, the expression for maximum time step is not precise, but a guideline of stability requirements.

8. Richards' equation in the water content form cannot solve variably saturated flow problems, only unsaturated flow problems.

9. Richards' equation in the mixed form cannot solve -per se- variably saturated flow problems. Appropriate discretization of the time derivative in order to obtain a discretized mixed form which solves for pressure is necessary.

10. Richards' equation in pressure form is apt to solve variably saturated flow problems, but is not sufficient to guarantee an accurate solution because of poor mass conservation properties.

11. The upstream mean generates faster wetting fronts than the arithmetic mean both for the EMC and IMC schemes.

12. The selection of the conductivity averaging technique is sensitive for coarse grids but less important in fine grids.

13. The effects of the conductivity averaging technique are insensitive to time step selection in the IMC scheme.

14. The use of the arithmetic mean together with IMC can generate overshooting solutions when near Dirichlet-type boundary conditions, that may result in failure to converge because of erroneously large pressure gradients.

15. The MMG model generates faster wetting fronts than the MG model. Drying fronts appear to be less sensitive.

16. It is more efficient from the computational perspective when using the IMC scheme to choose larger $\Delta t$ than selecting coarser grids. The use of coarse grids results in less accurate results with larger CPU times than when using large time steps.

## 4.2 Further research

1. Experimental validation is still necessary to verify all aspects of the model. In this work only validation against an analytical solution presented in the literature was performed, for a particular soil for which information is reported. Experiments would provide further insights into the effects of particular soil characteristics on the numerical response of the model. Furthermore, experimental benchmarking is scarce in the literature, and in most cases oriented to 2D saturated models, or 3D geotechnical models.

2. Substance transport coupled with variably saturated flow is a natural follow-up of this work. For this phenomenon, both the mathematical and numerical model are to be studied. The numerical effects over the schemes presented in this work are to be analyzed as well.

3. The analysis of the mathematical and numerical properties of Richards' equation and the numerical schemes performed in this work are oriented towards the develompent of a 3D variably saturated model which can interact with surface flow. The conclusions from this work are naturally put to use in the development of a 3D finite volume scheme which will interact with a 2D surface flow model in order to appropriately simulate interactions in hydrological systems such as rivers.

4. Applications to irrigation problems and inverse modeling techinques to obtain soil parameters are to be explored.

# Bibliography


[1] Anderson, D., Tannehill, J., and Pletcher, R. *Computational Fluid Mechanics and Heat Transfer*. Hemisphere Publishing Corporation, 1984.

[2] Baker, D. L. A Darcian integral approximation to interblock hydraulic conductivity means in vertical infiltration. *Computers & Geosciences 26*, 5 (2000), 581 – 590.

[3] Baker, D. L. General validity of conductivity means in unsaturated flow models. *Journal of Hydrologic Engineering 11*, 6 (NOV-DEC 2006), 526–538.

[4] Bear, J. *Dynamics of Fluids in Porous Media*. Dover Publications, 1988.

[5] Bear, J., and Bachmat, Y. *Introduction to Modeling of Transport Phenomena in Porous Media*. Kluwer Academic Publishers, 1990.

[6] Bear, J., and Verruijt, A. *Modeling Groundwater Flow and Pollution*. D. Reidel Publishing Company, 1998.

[7] Bedient, P., Huber, W., and Vieux, B. *Hydrology and Floodplain Analysis*. Prentice-Hall, 2008.

[8] Belfort, B., and Lehmann, F. Comparison of equivalent conductivities for numerical simulation of one-dimensional unsaturated flow. *Vadose Zone Journal 4*, 4 (NOV 2005), 1191–1200.

[9] Brooks, R., and Corey, A. Hydraulic properties of porous media, hydraulic paper # 3. Colorado State University, Fort Collins, CO, 1964.

[10] Brunone, B., Ferrante, M., Romano, N., and Santini, A. Numerical simulations of one-dimensional infiltration into layered soils with the Richards equation using diferrent estimates of the interlayer ocnductivity. *Vadose Zone Journal 2*, 2 (MAY 2003), 193–200.

[11] Brutsaert, W. *Hydrology*. Cambridge University Press, 2005.

[12] Børgesen, C. D., Jacobsen, O. H., Hansen, S., and Schaap, M. G. Soil hydraulic properties near saturation, an improved conductivity model. *Journal of Hydrology 324*, 1-4 (2006), 40 – 50.

[13] Celia, M., Bouloutas, E., and Zarba, R. A General Mass-Conservative Numerical-Solution for the Unsaturated Flow Equation. *Water Resources Research 26*, 7 (JUL 1990), 1483–1496.

[14] Desbarats, A. An interblock conductivity scheme for finite-difference models of steady unsaturated flow in heterogeneous media. *Water Resources Research 31*, 11 (NOV 1995), 2883–2889.

[15] Forsyth, P. A., Wu, Y. S., and Pruess, K. Robust numerical methods for saturated-unsaturated flow with dry initial conditions in heterogeneous media. *Advances in Water Resources 18*, 1 (1995), 25 – 38.

[16] Gardner, W. Some steady-state solutions of the unsaturated moisture flow equation with application to evaporation from a water table. *Soil Society 85* (1958), 228–232.

[17] Gastó, J., Grifoll, J., and Cohen, Y. Estimation of internodal permeabilities for numerical simulation of unsaturated flows. *Water Resources Research 38*, 12 (DEC 31 2002).

[18] Guarracino, L., and Basombrío, F. Un esquema no iterativo de segundo orden para la aproximación temporal de la ecuación no lineal de richards. *Mecánica computacional 23* (NOV 2004), 3059 – 3068.

[19] Gupta, R. *Hydrology and Hydraulic Systems*. Waveland Press, 2008.

[20] Haverkamp, R., and Vauclin, M. Note on estimating finite-difference interblock hydraulic conductivity values for transient unsaturated flow problems. *Water Resources Research 15*, 1 (1979), 181–187.

[21] Hornung, U., and Messing, W. Truncation errors in the numerical solution of horizontal diffusion in saturated/unsaturated media. *Advances in Water Resources 6*, 3 (1983), 165 – 168.

[22] Huang, K., Mohanty, B. P., and van Genuchten, M. T. A new convergence criterion for the modified Picard iteration method to solve the variably saturated flow equation. *Journal of Hydrology 178*, 1-4 (1996), 69 – 91.

[23] Jury, W., and Horton, R. *Soil Physics*. John Wiley and Sons, 2004.

[24] Kolditz, O. *Computational methods in environmental fluid mechanics*. Springer, 2002.

[25] Lehmann, F., and Ackerer, P. Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media. *Transport in Porous Media 31*, 3 (JUN 1998), 275–292.

[26] Paniconi, C., and Putti, M. A comparison of Picard and Newton iteration in the numerical solution of multidimensional variably satruated flow problems. *Water Resources Research 30*, 12 (DEC 1994), 3357–3374.

[27] Pei, Y., Wang, J., Tian, Z., and Yu, J. Analysis of interfacial error in saturated-unsaturated flow models. *Advances in Water Resources 29*, 4 (2006), 515 – 524.

[28] Phoon, K.-K., Tan, T.-S., and Chong, P.-C. Numerical simulation of Richards equation in partially saturated porous media: under-relaxation and mass balance. *Geotechnical and Geological Engineering 25*, 5 (OCT 2007), 525–541.

[29] Rathfelder, K., and Abriola, L. Mass conservative numerical-solutions of the head-based Richards equation. *Water Resources Research 30*, 9 (SEP 1994), 2579–2586.

[30] Richards, L. A. Capillary conduction of liquids through porous mediums. *Physics 1*, 5 (1931), 318–333.

[31] Romano, N., Brunone, B., and Santini, A. Numerical analysis of one-dimensional unsaturated flow in layered soils. *Advances in Water Resources 21*, 4 (1998), 315 – 324.

[32] Russo, D. Determining soil hydraulic-properties by parameter-estimation - on the selection of a model for the hydraulic-properties. *Water Resources Research 24*, 3 (MAR 1988), 453–459.

[33] Schaap, M., and van Genuchten, M. A modified Mualem-van Genuchten formulation for improved description of the hydraulic conductivity near saturation. *Vadose Zone Journal 5*, 1 (FEB 2006), 27–34.

[34] Sillers, W., Fredlund, D., and Zakerzaheh, N. Mathematical attributes of some soil water characteristic curve models. *Geotechnical and Geological Engineering 19*, 3 (2001), 243–283.

[35] Srivastava, R., and Guzman-Guzman, A. Analysis of hydraulic conductivity averaging schemes for one-dimensional, steady-state unsaturated flow. *Ground Water 33*, 6 (NOV-DEC 1995), 946–952.

[36] van Dam, J. C., and Feddes, R. A. Numerical simulation of infiltration, evaporation and shallow groundwater levels with the Richards equation. *Journal of Hydrology 233*, 1-4 (2000), 72 – 85.

[37] van Genuchten, M. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal 44*, 5 (1980), 892–898.

[38] Vanderborght, J., Kasteel, R., Herbst, M., Javaux, M., Thiery, D., Vanclooster, M., Mouvet, C., and Vereecken, H. A set of analytical benchmarks to test numerical models of flow and transport in soils. *Vadose Zone Journal 4*, 1 (FEB 2005), 206–221.

[39] Vogel, T., van Genuchten, M., and Cislerova, M. Effect of the shape of the soil hydraulic functions near saturation on variably-saturated flow predictions. *Advances in Water Resources 24*, 2 (NOV 2000), 133–144.

[40] Warrick, A. Numerical approximations of Darcian flow through unsaturated soil. *Water Resources Research 27*, 6 (JUN 1991), 1215–1222.

[41] Warrick, A., Lomen, D., and Yates, S. A generalized solution to infiltration. *Soil Science Society of America Journal 49*, 1 (1985), 34–38.

[42] Zaidel, J., and Russo, D. Estimation of Finite-Difference interblock conductivities for simulation of infiltration into initially dry soils. *Water Resources Research 28*, 9 (SEP 1992), 2285–2295.

newpage

# Appendix A

# Numerical schemes formulations

## A.1 Explicit Mixed Scheme

From Richards' equation in the mixed form, with a forward Euler scheme for time derivatives and a centered finite-difference scheme for spatial derivatives,

$$\frac{\theta^{n+1}-\theta^{n}}{\Delta t}-\frac{\partial}{\partial z}\left(K(h^{n})\frac{\partial h^{n}}{\partial z}+K(h^{n})\right)=0$$

$$\frac{\theta_i^{n+1}-\theta_i^{n}}{\Delta t}=\frac{K_{i+1/2}^{n}\frac{\partial h^{n}}{\partial z}\big|_{i+1/2}+K_{i+1/2}^{n}-K_{i-1/2}^{n}\frac{\partial h^{n}}{\partial z}\big|_{i-1/2}-K_{i-1/2}^{n}}{\delta z}$$

$$\frac{\theta_i^{n+1}-\theta_i^{n}}{\Delta t}=\frac{1}{\delta z}\left[K_{i+1/2}^{n}\frac{h_{i+1}^{n}-h_i^{n}}{\delta z}-K_{i-1/2}^{n}\frac{h_i^{n}-h_{i-1}^{n}}{\delta z}+K_{i+1/2}^{n}-K_{i-1/2}^{n}\right]$$

$$\frac{\theta_i^{n+1}-\theta_i^{n}}{\Delta t}=\frac{1}{\delta z}\left[K_{i+1/2}^{n}\frac{h_{i+1}^{n}}{\delta z}-\left(K_{i+1/2}^{n}+K_{i-1/2}^{n}\right)\frac{h_i^{n}}{\delta z}+K_{i-1/2}^{n}\frac{h_{i-1}^{n}}{\delta z}+\left(K_{i+1/2}^{n}-K_{i-1/2}^{n}\right)\right]$$

$$\theta_i^{n+1}=\frac{\Delta t}{\delta z^{2}}K_{i+1/2}^{n}h_{i+1}^{n}-\frac{\Delta t}{\delta z^{2}}\left(K_{i+1/2}^{n}+K_{i-1/2}^{n}\right)h_i^{n}+\frac{\Delta t}{\delta z^{2}}K_{i-1/2}^{n}h_{i-1}^{n}+\frac{\Delta t}{\delta z}\left(K_{i+1/2}^{n}-K_{i-1/2}^{n}\right)+\theta_i^{n} \tag{A.1}$$

## A.2 Explicit pressure based scheme

From Richards' equation in pressure form, with a forward Euler scheme for time derivatives and a centered finite-difference scheme for spatial derivatives,

$$C^{n}\frac{h^{n+1}-h^{n}}{\Delta t}-\frac{\partial}{\partial z}\left(K(h^{n})\frac{\partial h^{n}}{\partial z}+K(h^{n})\right)=0$$

$$C_i^{n}\frac{h_i^{n+1}-h_i^{n}}{\Delta t}=\frac{K_{i+1/2}^{n}\frac{\partial h^{n}}{\partial z}\big|_{i+1/2}+K_{i+1/2}^{n}-K_{i-1/2}^{n}\frac{\partial h^{n}}{\partial z}\big|_{i-1/2}-K_{i-1/2}^{n}}{\delta z}$$

$$C_i^n \frac{h_i^{n+1} - h_i^n}{\Delta t} = \frac{1}{\delta z}\left[K_{i+1/2}^n \frac{h_{i+1}^n - h_i^n}{\delta z} - K_{i-1/2}^n \frac{h_i^n - h_{i-1}^n}{\delta z} + K_{i+1/2}^n - K_{i-1/2}^n\right]$$

$$h_i^{n+1} = \frac{\Delta t}{C_i^n \delta z^2} K_{i+1/2}^n h_{i+1}^n - \frac{\Delta t}{C_i^n \delta z^2}\left(K_{i+1/2}^n + K_{i-1/2}^n\right) h_i^n + \frac{\Delta t}{C_i^n \delta z^2} K_{i-1/2}^n h_{i-1}^n + \frac{\Delta t}{C_i^n \delta z}\left(K_{i+1/2}^n - K_{i-1/2}^n\right) + h_i^n \quad \text{(A.2)}$$

## A.3 Implicit Presure based scheme

From Richards' equation in pressure form, with a backward Euler scheme for time derivatives and a centered finite-difference scheme for spatial derivatives,

$$C^{n+1} \frac{h^{n+1} - h^n}{\Delta t} - \frac{\partial}{\partial z}\left[K(h^{n+1})\left(\frac{\partial h^{n+1}}{\partial z} + 1\right)\right] = 0$$

In order to formulate the Picard iteration, let $\delta^m = h^{n+1,m+1} - h^{n+1,m}$ where $m+1$ is the computed iteration and $m$ the previous iteration.

$$C^{n+1,m}\frac{\delta^m}{\Delta t} + C^{n+1,m}\frac{h^{n+1,m} - h^n}{\Delta t} - \frac{\partial}{\partial z}\left(K^{n+1,m}\frac{\partial \delta^m}{\partial z}\right) = \frac{\partial K^{n+1,m}}{\partial z} + \frac{\partial}{\partial z}\left(K^{n+1,m}\frac{\partial h^{n+1,m}}{\partial z}\right)$$

Spatial discretization with finite differences, using $\frac{\partial \delta^m}{\partial z}\big|_{i\pm1/2} \approx \frac{\delta_i^m - \delta m_{i\pm1}}{\delta z}$ and evaluating hydraulic conductivity in between cells yields

$$C_i^{n+1,m}\frac{\delta_i^m}{\Delta t} + C_i^{n+1,m}\frac{h_i^{n+1,m} - h_i^n}{\Delta t} - \frac{1}{\delta z^2}\left[K_{i+1/2}^{n+1,m}\left(\delta_{i+1}^m - \delta_i^m\right) - K_{i-1/2}^{n+1,m}\left(\delta_i^m - \delta_{i-1}^m\right)\right] = \frac{1}{\delta z}\left(K_{i+1/2}^{n+1,m} - K_{i-1/2}^{n+1,m}\right) + \frac{1}{\delta z^2}\left[K_{i+1/2}^{n+1,m}\left(h_{i+1}^{n+1,m} - h_i^{n+1,m}\right) - K_{i-1/2}^{n+1,m}\left(h_i^{n+1,m} - h_{i-1}^{n+1,m}\right)\right]$$

Expanding all terms,

$$C_i^{n+1,m}\frac{\delta_i^m}{\Delta t} + C_i^{n+1,m}\frac{h_i^{n+1,m} - h_i^n}{\Delta t} - \frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}\delta_{i+1}^m + \frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}\delta_i^m + \frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}\delta_i^m - \frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}\delta_{i-1}^m = \frac{1}{\delta z}\left(K_{i+1/2}^{n+1,m} - K_{i-1/2}^{n+1,m}\right) + \frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_{i+1}^{n+1,m} - \frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_i^{n+1,m} - \frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_i^{n+1,m} + \frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_{i-1}^{n+1,m}$$

Grouping terms by spatial index $i-1$, $i$, and $i+1$

$$-\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}\delta_{i-1}^m - \frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_{i-1}^{n+1,m} + \frac{C_i^{n+1,m}\delta_i^m}{\Delta t} + C_i^{n+1,m}\frac{h_i^{n+1,m}}{\Delta t} - C_i^{n+1,m}\frac{h_i^n}{\Delta t} + \frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}\delta_i^m + \frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}\delta_i^m + \frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_i^{n+1,m} + \frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_i^{n+1,m} - \frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}\delta_{i+1}^m - \frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_{i+1}^{n+1,m} = \frac{1}{\delta z}\left(K_{i+1/2}^{n+1,m} - K_{i-1/2}^{n+1,m}\right)$$

Expanding $\delta^m$ with its definition,

$$
\begin{aligned}
&-\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_{i-1}^{n+1,m+1}+\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_{i-1}^{n+1,m}-\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_{i-1}^{n+1,m}\\
&\quad+\frac{C_i^{n+1,m}h_i^{n+1,m+1}}{\Delta t}-\frac{C_i^{n+1,m}h_i^{n+1,m}}{\Delta t}+C_i^{n+1,m}\frac{h_i^{n+1,m}}{\Delta t}-C_i^{n+1,m}\frac{h_i^{n}}{\Delta t}\\
&\quad+\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_{i}^{n+1,m+1}-\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_{i}^{n+1,m}\\
&+\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_{i}^{n+1,m+1}-\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_{i}^{n+1,m}+\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_{i}^{n+1,m}+\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_{i}^{n+1,m}\\
&\quad-\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_{i+1}^{n+1,m+1}+\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_{i+1}^{n+1,m}-\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_{i+1}^{n+1,m}=\\
&\qquad\qquad\frac{1}{\delta z}\left(K_{i+1/2}^{n+1,m}-K_{i-1/2}^{n+1,m}\right)
\end{aligned}
$$

Grouping and rearranging, results in

$$
\left(-\frac{\Delta t}{\delta z^2}K_{i-1/2}^{n+1,m}\right)h_{i-1}^{n+1,m+1}+\left[C_i^{n+1,m}+\frac{\Delta t}{\delta z^2}\left(K_{i+1/2}^{n+1,m}+K_{i-1/2}^{n+1,m}\right)\right]h_i^{n+1,m+1}+\left(-\frac{\Delta t}{\delta z^2}K_{i+1/2}^{n+1,m}\right)h_{i+1}^{n+1,m+1}=
$$
$$
\frac{\Delta t}{\delta z}\left(K_{i+1/2}^{n+1,m}-K_{i-1/2}^{n+1,m}\right)+C_i^{n+1,m}h_i^{n}
$$

## A.4 Implicit Mixed Conservative scheme

From Richards' equation in the mixed form, with a backward Euler scheme for time derivatives and a centered finite-difference scheme for spatial derivatives,

$$
\frac{\theta^{n+1}-\theta^n}{\Delta t}-\frac{\partial}{\partial z}\left[K(h^{n+1})\left(\frac{\partial h^{n+1}}{\partial z}+1\right)\right]=0
$$

In order to formulate the Picard iteration, let $\delta^m=h^{n+1,m+1}-h^{n+1,m}$ where $m+1$ is the computed iteration and $m$ the previous iteration. Invoking Taylor's polynomial to approximate $\theta^{n+1,m+1}$ in terms of the derivative $C=\dfrac{\partial\theta}{\partial h}$ and $\delta^m$,

$$
\frac{\theta^{n+1,m}+C^{n+1,m}\delta^m-\theta^n}{\Delta t}-\frac{\partial}{\partial z}\left(K^{n+1,m}\frac{\partial\delta^m}{\partial z}\right)=\frac{\partial K^{n+1,m}}{\partial z}+\frac{\partial}{\partial z}\left(K^{n+1,m}\frac{\partial h^{n+1,m}}{\partial z}\right)
$$

Spatial discretization with finite differences, using $\dfrac{\partial\delta^m}{\partial z}\Big|_{i\pm1/2}\approx\dfrac{\delta_i^m-\delta m_{i\pm1}}{\delta z}$ and evaluating hydraulic conductivity in between cells yields

$$
\begin{aligned}
&\frac{\theta_i^{n+1,m}+C_i^{n+1,m}\delta_i^m-\theta_i^n}{\Delta t}-\frac{1}{\delta z^2}\left[K_{i+1/2}^{n+1,m}\left(\delta_{i+1}^m-\delta_i^m\right)-K_{i-1/2}^{n+1,m}\left(\delta_i^m-\delta_{i-1}^m\right)\right]=\\
&\frac{1}{\delta z}\left(K_{i+1/2}^{n+1,m}-K_{i-1/2}^{n+1,m}\right)+\frac{1}{\delta z^2}\left[K_{i+1/2}^{n+1,m}\left(h_{i+1}^{n+1,m}-h_i^{n+1,m}\right)-K_{i-1/2}^{n+1,m}\left(h_i^{n+1,m}-h_{i-1}^{n+1,m}\right)\right]
\end{aligned}
$$

Expanding all terms,

$$
\begin{aligned}
&\frac{\theta_i^{n+1,m}-\theta_i^n}{\Delta t}+\frac{C_i^{n+1,m}\delta_i^m}{\Delta t}-\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}\delta_{i+1}^m+\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}\delta_i^m+\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}\delta_i^m-\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}\delta_{i-1}^m=\\
&\frac{1}{\delta z}\left(K_{i+1/2}^{n+1,m}-K_{i-1/2}^{n+1,m}\right)+\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_{i+1}^{n+1,m}-\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_i^{n+1,m}-\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_i^{n+1,m}+\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_{i-1}^{n+1,m}
\end{aligned}
$$

Grouping terms by spatial index $i-1$, $i$, and $i+1$

$$
\begin{aligned}
&-\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}\delta_{i-1}^{m}-\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_{i-1}^{n+1,m}\\
&+\frac{C_i^{n+1,m}\delta_i^m}{\Delta t}+\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}\delta_i^m+\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}\delta_i^m+\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_i^{n+1,m}+\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_i^{n+1,m}\\
&-\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}\delta_{i+1}^{m}-\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_{i+1}^{n+1,m}=\\
&\frac{1}{\delta z}\left(K_{i+1/2}^{n+1,m}-K_{i-1/2}^{n+1,m}\right)-\frac{\theta_i^{n+1,m}-\theta_i^n}{\Delta t}
\end{aligned}
$$

Expanding $\delta^m$ with its definition,

$$
\begin{aligned}
&-\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_{i-1}^{n+1,m+1}+\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_{i-1}^{n+1,m}-\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_{i-1}^{n+1,m}\\
&+\frac{C_i^{n+1,m}h_i^{n+1,m+1}}{\Delta t}-\frac{C_i^{n+1,m}h_i^{n+1,m}}{\Delta t}+\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_i^{n+1,m+1}-\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_i^{n+1,m}\\
&+\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_i^{n+1,m+1}-\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_i^{n+1,m}+\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_i^{n+1,m}+\frac{1}{\delta z^2}K_{i-1/2}^{n+1,m}h_i^{n+1,m}\\
&-\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_{i+1}^{n+1,m+1}+\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_{i+1}^{n+1,m}-\frac{1}{\delta z^2}K_{i+1/2}^{n+1,m}h_{i+1}^{n+1,m}=\\
&\frac{1}{\delta z}\left(K_{i+1/2}^{n+1,m}-K_{i-1/2}^{n+1,m}\right)-\frac{\theta_i^{n+1,m}-\theta_i^n}{\Delta t}
\end{aligned}
$$

Grouping and rearranging, results in

$$
\begin{aligned}
&\left(-\frac{\Delta t}{\delta z^2}K_{i-1/2}^{n+1,m}\right)h_{i-1}^{n+1,m+1}+\left[C_i^{n+1,m}+\frac{\Delta t}{\delta z^2}\left(K_{i+1/2}^{n+1,m}+K_{i-1/2}^{n+1,m}\right)\right]h_i^{n+1,m+1}+\left(-\frac{\Delta t}{\delta z^2}K_{i+1/2}^{n+1,m}\right)h_{i+1}^{n+1,m+1}=\\
&C_i^{n+1,m}h_i^{n+1,m}+\frac{\Delta t}{\delta z}\left(K_{i+1/2}^{n+1,m}-K_{i-1/2}^{n+1,m}\right)+\theta_i^n-\theta_i^{n+1,m}
\end{aligned}
$$

# Appendix B

# Stability Analysis

## B.1 EMC Scheme

Recall equation (1.40), the mixed form of Richards' equation:

$$\frac{\partial \theta(h)}{\partial t} = \frac{\partial}{\partial z}\left[K(h)\left(\frac{\partial h}{\partial z} + 1\right)\right] = -\frac{\partial J}{\partial z}$$

In order to perform the stability analysis consider the following:

- Conductivity is assumed as linear function of water content: $K(\theta) = K_o + K_*(\theta - \theta_o)$
- Water content is assumed as linear function of water head: $\theta(h) = \theta_o + C(h - h_o)$.

Consider a soil column with an initially uniform moisture distribution in the entire depth. Let $\tilde{\theta}$ be a perturbation of the initial moisture distribution, described by

$$\tilde{\theta} = a + be^{\mathbf{i}\psi z} \tag{B.1}$$

In consequence,

$$\frac{\partial \tilde{\theta}}{\partial z} = \frac{\partial}{\partial z}\left(a + be^{\mathbf{i}\psi z}\right) = \mathbf{i}b\psi e^{\mathbf{i}\psi z}$$

Because $\theta = \theta(h)$

$$\frac{\partial \tilde{\theta}}{\partial h}\frac{\partial h}{\partial z} = C\frac{\partial h}{\partial z} = \mathbf{i}b\psi e^{\mathbf{i}\psi z}$$

Rearranging,

$$\frac{\partial h}{\partial z} = \frac{\mathbf{i}b\psi}{C}e^{\mathbf{i}\psi z} \tag{B.2}$$

From the definition of the flux $J$, and considering equation (B.2)

$$J = -K\left(\frac{\partial h}{\partial z} + 1\right) = -K\frac{\mathbf{i}b\psi}{C}e^{\mathbf{i}\psi z} - K$$

Introducing the linear form of $K$,

$$J = -\left(K_o + K_*\theta - K_*\theta_o\right)\left(\frac{\mathbf{i}b\psi}{C}e^{\mathbf{i}\psi z} + 1\right)$$

Hence, flux $\tilde{J}$ for the perturbation $\theta = \tilde{\theta}$ is

$$\tilde{J} = -\left(K_o - K_*\theta_o\right)\left(\frac{\mathbf{i}b\psi}{C}e^{\mathbf{i}\psi z} + 1\right) - K_*\left(a + be^{\mathbf{i}\psi z}\right)\left(\frac{\mathbf{i}b\psi}{C}e^{\mathbf{i}\psi z} + 1\right)$$

Regrouping,

$$\tilde{J} = -K_o + K_*\theta_o - K_*a - \left(K_o - K_*\theta_o + K_*a + K_*b\right)\left(\frac{\mathbf{i}b\psi}{C}e^{\mathbf{i}\psi z}\right) - K_*\frac{\mathbf{i}b^2\psi}{C}e^{2\mathbf{i}\psi z}$$

Because $\theta_o = a$ for perturbation conditions,

$$\tilde{J} = -K_o \underbrace{-\left(K_o + K_*b\right)\left(\frac{\mathbf{i}b\psi}{C}e^{\mathbf{i}\psi z}\right)}_{\tilde{J}_1} \underbrace{-K_*\frac{\mathbf{i}b^2\psi}{C}e^{2\mathbf{i}\psi z}}_{\tilde{J}_2} \tag{B.3}$$

A first order Taylor expansion of $\tilde{J}(\xi)$ in the neigbourhood of $\xi = \varphi$ is

$$\tilde{J}(\xi) = \tilde{J}(\varphi) + \frac{\partial\tilde{J}(\varphi)}{\partial\xi}\left(\xi - \varphi\right) + O(\xi^2)$$

Consider $\xi = e^{\mathbf{i}\psi z}$ and $\varphi = \xi(z = 0) = 1$. The derivative required for the Taylor expansion is

$$\frac{\partial\tilde{J}}{\partial\xi} = \frac{\partial\tilde{J}_1(\xi)}{\partial\xi} + \frac{\partial\tilde{J}_2(\xi)}{\partial\xi}$$

where

$$\frac{\partial\tilde{J}_1(\xi)}{\partial\xi} = -\left(K_o + K_*b\right)\left(\frac{\mathbf{i}b\psi}{C}\right)$$

and

$$\frac{\partial\tilde{J}_2(\xi)}{\partial\xi} = -2K_*\frac{\mathbf{i}b^2\psi}{C}\xi$$

Hence,

$$\frac{\partial\tilde{J}}{\partial\xi} = -\left(K_o + K_*b\right)\left(\frac{\mathbf{i}b\psi}{C}\right) - 2K_*\frac{\mathbf{i}b^2\psi}{C}\xi$$

In consequence, the Taylor polynomial, neglecting $O(\xi^2)$, is evaluated as

$$\tilde{J} = -K_o - \left(K_o + K_*b\right)\left(\frac{\mathbf{i}b\psi}{C}\right) - 2K_*\frac{\mathbf{i}b^2\psi}{C} - \left[\left(K_o + K_*b\right)\left(\frac{\mathbf{i}b\psi}{C}\right) + 2K_*\frac{\mathbf{i}b^2\psi}{C}\right]\left(e^{\mathbf{i}\psi z} - 1\right)$$

Finally, the perturabated flux is

$$\tilde{J} = -K_o - \left(K_o + 3K_*b\right)\frac{\mathbf{i}b\psi}{C}e^{\mathbf{i}\psi z} \tag{B.4}$$

Consider the EMC scheme, written in terms of flux $J$,

$$\theta_i^{n+1} = \theta_i^n - \frac{\Delta t}{\delta z}\left(J_{i+1/2}^n - J_{i-1/2}^n\right) \tag{B.5}$$

Consider that in time $n$ the perturbation $\tilde{\theta}$ and the perturbation flow $\tilde{J}$ occur, hence $\theta^n = \tilde{\theta}$ and $J^n = \tilde{J}$. Hence, $J^n_{i\pm1/2} = \tilde{J}(z \pm \delta z)$. Consequently, by substituting equations (B.1) and (B.4) into equation (B.5) yields

$$\theta_i^{n+1} = a + be^{\mathbf{i}\psi z_i} + \frac{\Delta t}{\delta z}\left[K_o + (K_o + 3K_*b)\,\frac{\mathbf{i}b\psi}{C}e^{\mathbf{i}\psi(z+\delta z_i/2)}\right] - \frac{\Delta t}{\delta z}\left[K_o + (K_o + 3K_*b)\,\frac{\mathbf{i}b\psi}{C}e^{\mathbf{i}\psi(z-\delta z_i/2)}\right]$$

Cancelling terms and rearranging,

$$\theta_i^{n+1} = a + be^{\mathbf{i}\psi z_i} + \frac{\Delta t}{\delta z}\,(K_o + 3K_*b)\,\frac{\mathbf{i}b\psi}{C}e^{\mathbf{i}\psi z}e^{\mathbf{i}\psi\delta z/2} - \frac{\Delta t}{\delta z}\,(K_o + 3K_*b)\,\frac{\mathbf{i}b\psi}{C}e^{\mathbf{i}\psi z}e^{-\mathbf{i}\psi\delta z/2}$$

Grouping,

$$\theta_i^{n+1} = a + be^{\mathbf{i}\psi z_i}\left\{1 + \frac{\Delta t}{\delta z}\,(K_o + 3K_*b)\,\frac{\mathbf{i}\psi}{C}\left(e^{\mathbf{i}\psi\delta z/2} - e^{-\mathbf{i}\psi\delta z/2}\right)\right\} \tag{B.6}$$

Equation (B.6) is the expression which describes how a perturbation in an an initially uniform water content profile progresses in time. In order for the scheme to be stable, the perturbation cannot amplify itself. Hence,

$$\left|1 + \frac{\Delta t}{\delta z}\,(K_o + 3K_*b)\,\frac{\mathbf{i}\psi}{C}\left(e^{\mathbf{i}\psi\delta z/2} - e^{-\mathbf{i}\psi\delta z/2}\right)\right|^2 \leq 1$$

Because of the relation of the complex exponential function to tringonometric functions,

$$\left|1 + \frac{\Delta t}{\delta z}\,(K_o + 3K_*b)\,\frac{\mathbf{i}\psi}{C}\left(2\mathbf{i}\sin\left(\frac{\psi\delta z}{2}\right)\right)\right|^1 \leq 1$$

Hence,

$$\left|1 - \frac{2\psi}{C}\frac{\Delta t}{\delta z}\,(K_o + 3K_*b)\sin\left(\frac{\psi\delta z}{2}\right)\right|^2 \leq 1$$

Consider that the smallest wave length that can be observed in a uniform space discretization of size $\delta z$ is $\lambda = 4\delta z$, which implies that the the wave number satisifies $\psi = \dfrac{\pi}{2\delta z}$, which implies $\sin\left(\dfrac{\psi\delta}{2}\right) = \sin\left(\dfrac{\pi}{4}\right) = \dfrac{\sqrt{2}}{2}$. Hence,

$$\left|1 - \pi\frac{\Delta t}{C\delta z^2}\Big(K_o + 3K_*b\Big)\right|^2 \leq 1$$

Expanding the squared modulus,

$$1 + \pi^2\frac{\Delta t^2}{C^2\delta z^4}\Big(K_o + 3K_*b\Big)^2 - 2\pi\frac{\Delta t}{C\delta z^2}\Big(K_o + 3K_*b\Big) \leq 1$$

Solving for $\Delta t$,

$$\Delta t \leq \frac{2}{\pi}\frac{\delta z^2 C}{K_o + 3K_*b}$$

Consider the definition of a dimensionless number $\epsilon_*$,

$$\epsilon_* = \frac{3K_* b}{K_o} \tag{B.7}$$

And consider the following defintion of parameter $\nu_*$ with viscosity units,

$$\nu_* = \frac{K_o}{C} \tag{B.8}$$

Then, the stability criterion is

$$\Delta t \leq \frac{2}{\pi}\frac{\delta z^2}{\nu_*}\left(\frac{1}{1+\epsilon_*}\right) \tag{B.9}$$

Consider equation (B.9), where $\epsilon_* \ll 1$, then

$$\Delta t \leq \frac{2}{\pi}\frac{\delta z^2}{\nu_*}$$

Note that $\nu_*$ has viscosity units $\frac{L^2}{T}$ and that the stability criterion resembles the well-known stability criterion for an explicit centered finite difference scheme for the 1D diffusion equation $\Delta t \leq \frac{\delta x^2}{2\alpha}$ where $\alpha$ is a constant diffusion coefficient [1]. However, $\nu_*$ is not a constant coefficient, and the stability criterion depends on $K_o$ and $C$.

**Effects of $\nu_*$ and $\epsilon_*$**

To further investigate stability, consider the Brooks-Corey model [9]:

$$\theta = \left(\frac{h_b}{h}\right)^{\omega}(\theta_s - \theta_r) + \theta_r \tag{B.10}$$

where $h_b$ is the *bubbling pressure* and $\omega$ a fitting parameter that represents pore-size distribution. Hence, its derivative is

$$C = \frac{\partial\theta}{\partial h} = -\omega h_b^{\omega}\left(\theta_s - \theta_r\right)h^{-(\omega+1)} \tag{B.11}$$

Consider the conductivity function as $K = K_s\left(\frac{\theta-\theta_r}{\theta_s-\theta_r}\right)^{2+\frac{5}{2\omega}}$ as suggested by Brutsaert [11]. Then,

$$K_* = \frac{\partial K}{\partial\theta} = \frac{\left(2+\frac{5}{2\omega}\right)K_s}{(\theta_s-\theta_r)^{\frac{5}{2\omega}}}(\theta-\theta_r)^{\frac{5}{2\omega}+1} \tag{B.12}$$

Hence,

$$\nu_* = \frac{K_o}{C} = -\frac{K_s\left(\frac{\theta_o-\theta_r}{\theta_s-\theta_r}\right)^{2+\frac{5}{2\omega}}}{\omega h_b^{\omega}\left(\theta_s-\theta_r\right)h^{-(\omega+1)}}$$

Which is

$$\nu_* = -\frac{K_s}{\omega}h_b\frac{\theta_o-\theta_r}{(\theta_s-\theta_r)^2}\left(\frac{\theta_o-\theta_r}{\theta_s-\theta_r}\right)^{\frac{3}{2\omega}}$$

On the other hand,

$$\epsilon_* = \frac{3K_* b}{K_o} = 3b \frac{\frac{\left(2+\frac{5}{2\omega}\right) K_s}{(\theta_s-\theta_r)^{\frac{5}{2\omega}}} (\theta-\theta_r)^{\frac{5}{2\omega}+1}}{K_s \left(\frac{\theta-\theta_r}{\theta_s-\theta_r}\right)^{2+\frac{5}{2\omega}}}$$

Rearranging,

$$\epsilon_* = 3b \left(2+\frac{5}{2\omega}\right) \frac{(\theta_s-\theta_r)^2}{\theta_o-\theta_r}$$

Finally, the stability criterion can be expressed as

$$\Delta t = -\frac{2\delta z^2}{\pi} \frac{\omega(\theta_s-\theta_r)^2}{K_s h_b (\theta_o-\theta_r)} \left(\frac{\theta_s-\theta_r}{\theta_o-\theta_r}\right)^{\frac{3}{2\omega}} \left[\frac{1}{1+3b\left(2+\frac{5}{2\omega}\right)\frac{(\theta_s-\theta_r)^2}{\theta_o-\theta_r}}\right]$$

From this expression it can be concluded that, for a particular soil, $\Delta t$ is inversely proportional to saturation. The higher $\omega$ the more sensible $\Delta t$ is to saturation. Conversely, for a particular humidity, there is a minimum value of $\Delta t$ for a particular $\omega$. The more saturated the soil is, the least sensitive $\Delta t$ is to $\omega$. Saturated conditions result in a $\Delta t$ which varies with $\omega$ very little around the minimum value of $\Delta t$.

## B.2 IMC Scheme

Consider the perturbation of an initial state

$$\tilde{h} = a + be^{\mathbf{i}(\psi z - \omega t)} \tag{B.13}$$

Hence,

$$h_i^{n+1} = a + be^{\mathbf{i}\left(\psi z_i - \omega t^{n+1}\right)}$$

Assume that the perturbation does not suffer changes in wave length in space, hence:

$$h_{i+1}^{n+1} = a + be^{\mathbf{i}\left(\psi(z_i+\delta z) - \omega t^{n+1}\right)} = a + be^{\mathbf{i}\left(\psi z_i - \omega t^{n+1}\right)} e^{\mathbf{i}(\psi\delta z)} = a - ae^{\mathbf{i}\psi\delta z} + h_i^{n+1} e^{\mathbf{i}\psi\delta z}$$

In summary,

$$h_{i+1}^{n+1} = a - ae^{\mathbf{i}\psi\delta z} + h_i^{n+1} e^{\mathbf{i}\psi\delta z} \tag{B.14}$$

And in a similar way,

$$h_{i-1}^{n+1} = a - ae^{-\mathbf{i}\psi\delta z} + h_i^{n+1} e^{-\mathbf{i}\psi\delta z} \tag{B.15}$$

Let hydraulic conductivity and water content be linear functions of pressure, hence

$$K = K_o + K_h(h - h_o) \tag{B.16}$$

$$\theta = \theta_o + C(h - h_o) \tag{B.17}$$

Consider the IMC scheme as described by equation (2.15)

$$\left(-\frac{\Delta t}{\delta z^2}K_{i-1/2}^{n+1,m}\right)h_{i-1}^{n+1,m+1}+\left[C_i^{n+1,m}+\frac{\Delta t}{\delta z^2}\left(K_{i+1/2}^{n+1,m}+K_{i-1/2}^{n+1,m}\right)\right]h_i^{n+1,m+1}$$
$$+\left(-\frac{\Delta t}{\delta z^2}K_{i+1/2}^{n+1,m}\right)h_{i+1}^{n+1,m+1}=C_i^{n+1,m}h_i^{n+1,m}+\frac{\Delta t}{\delta z}\left(K_{i+1/2}^{n+1,m}-K_{i-1/2}^{n+1,m}\right)+\theta_i^n-\theta_i^{n+1,m}$$

Recall that iterations in this scheme respond mainly to mass-balance issues because of the linearization of $C$. Hence, for small amplitude perturbations of $h$ (and consequently of $C$ and $K$) consider time $m+1$ as $n+1$ for $h$ and time $m$ as time $n$ for $K$ and $C$. Note that for the scheme to converve in a single time step it is necessary that $\theta^m\approx\theta^{m+1}$. Hence, the scheme may be rewritten as

$$-\frac{\Delta t}{\delta z^2}K_{i-1/2}^n h_{i-1}^{n+1}+\left[C_i^n+\frac{\Delta t}{\delta z^2}\left(K_{i+1/2}^n+K_{i-1/2}^n\right)\right]h_i^{n+1}$$
$$-\frac{\Delta t}{\delta z^2}K_{i+1/2}^n h_{i+1}^{n+1}=C_i^n h_i^n+\frac{\Delta t}{\delta z}\left(K_{i+1/2}^n-K_{i-1/2}^n\right)+\theta_i^n-\theta_i^{n+1}\quad\text{(B.18)}$$

Substituting equations (B.14) and (B.15) in equation (B.18) and the linear definition of $\theta$,

$$-\frac{\Delta t}{\delta z^2}K_{i-1/2}^n\left(a-ae^{-\mathbf{i}\psi\delta z}+h_i^{n+1}e^{-\mathbf{i}\psi\delta z}\right)+\left[C_i^n+\frac{\Delta t}{\delta z^2}\left(K_{i+1/2}^n+K_{i-1/2}^n\right)\right]h_i^{n+1}$$
$$-\frac{\Delta t}{\delta z^2}K_{i+1/2}^n\left(a-ae^{\mathbf{i}\psi\delta z}+h_i^{n+1}e^{\mathbf{i}\psi\delta z}\right)=C_i^n h_i^n+\frac{\Delta t}{\delta z}\left(K_{i+1/2}^n-K_{i-1/2}^n\right)$$
$$+\theta_o+C_i^n(h_i^n-h_o)-\theta_o-C_i^n(h_i^{n+1}-h_o)$$

Grouping,

$$\frac{\Delta t}{\delta z^2}h_i^{n+1}\left[\frac{2\delta z^2}{\Delta t}C_i^n-K_{i-1/2}^n e^{-\mathbf{i}\psi\delta z}+K_{i+1/2}^n+K_{i-1/2}^n-K_{i+1/2}^n e^{\mathbf{i}\psi\delta z}\right]$$
$$-a\frac{\Delta t}{\delta z^2}\left[K_{i-1/2}^n\left(1-e^{-\mathbf{i}\psi\delta z}\right)+K_{i+1/2}^n\left(1-e^{\mathbf{i}\psi\delta z}\right)\right]=2C_i^n h_i^n+\frac{\Delta t}{\delta z}\left(K_{i+1/2}^n-K_{i-1/2}^n\right)\quad\text{(B.19)}$$

Because of the linear definition of $K$,

$$K_{i\pm1/2}=\frac{K_o+K_h(h_i-h_o)+K_o+K_h(h_{i\pm1}-h_o)}{2}=\frac{2K_o+K_h h_i+K_h h_{i\pm1}-2K_h h_o}{2}$$

$$K_{i\pm1/2}^n=K_o-K_h h_o+\frac{K_h}{2}\left(h_i^n+h_{i+\pm1}^n\right)$$

Furthermore, $h_{i\pm1}^n=a-ae^{\pm\mathbf{i}\psi\delta z}+h_i^n e^{\pm\mathbf{i}\psi\delta z}$. Hence,

$$K_{i\pm1/2}^n=K_o-K_h h_o+\frac{K_h}{2}\left(h_i^n+a-ae^{\pm\mathbf{i}\psi\delta z}+h_i^n e^{\pm\mathbf{i}\psi\delta z}\right)$$

Note that because of the perturbation analysis, $a=h_o$ and $h_i^n=h_o$, which results in

$$K_{i\pm1/2}^n=K_o$$

Let $C_i^n=C_h$. In consequence, equation (B.19) becomes

$$\frac{\Delta t}{\delta z^2}K_o h_i^{n+1}\left[\frac{2\delta z^2}{\Delta t}\frac{C_h}{K_o}+2-\left(e^{\mathbf{i}\psi\delta z}+e^{-\mathbf{i}\psi\delta z}\right)\right]-h_o K_o\frac{\Delta t}{\delta z^2}\left[2-\left(e^{\mathbf{i}\psi\delta z}+e^{-\mathbf{i}\psi\delta z}\right)\right]=2C_h h_o$$

By means of the identities between de complex exponential function and trigonometric functions,

$$\frac{\Delta t}{\delta z^2} K_o h_i^{n+1} \left[ \frac{\delta z^2}{\Delta t} \frac{C_h}{K_o} + 1 - \cos(\psi \delta z) \right] - h_o K_o \frac{\Delta t}{\delta z^2} \left[ 1 - \cos(\psi \delta z) \right] = C_h h_o$$

The amplification factor is $G = \dfrac{h_i^{n+1}}{h_o}$, hence

$$G \left[ \frac{\delta z^2}{\Delta t} \frac{C_h}{K_o} + 1 - \cos(\psi \delta z) \right] - 1 + \cos(\psi \delta z) = C_h \frac{\delta z^2}{K_o \Delta t}$$

Solving for $G$

$$G = \frac{1 + C_h \dfrac{\delta z^2}{K_o \Delta t} - \cos(\psi \delta z)}{1 + C_h \dfrac{\delta z^2}{K_o \Delta t} - \cos(\psi \delta z)} = 1 \tag{B.20}$$

This result implies that for small perturbations, the IMC method is unconditionally stable. If the analysis is performed for the first iteration, i.e., $\theta^{m+1} = \theta^n$ the same conclusion is obtained.